PARTS OF THE ORIGINAL DOCUMENT

ARE OF POOR REPRODUCTIVE QUALITY

CERTAINES PARTIES DU DOCUMENT ORIGINAL

SONT DE PIÈTRE QUALITÉ DE REPRODUCTION

Documents
de référence

Recueil des textes des présentations du colloque sur les applications de la statistique

75 ANS DE FIABILITÉ

Bureau de
la statistique
du Québec

Pour tout renseignement concernant le BSQ
et les données statistiques qui y sont disponibles,
s'adresser au:

Bureau de la statistique du Québec
117, rue Saint-André
Québec (Québec)
G1K 3Y3
Tél.: (418) 643-5116

ou

1410, rue Stanley, 3e étage
Montréal (Québec)
H3A 1P8
Tél.: (514) 873-4722

ou

Tél.: 1-800-463-4090
(aucun frais d'appel)

Dépôt légal
Bibliothèque nationale du Québec
Quatrième trimestre 1987
ISBN 2-550-18196-4

Tirage limité.
Toute reproduction ou utilisation du contenu
de cette publication est autorisée à condition
de mentionner la source.

Présentation

Ce recueil contient les textes des présentations du "Colloque sur les applications de la statistique" qui s'est tenu à Ottawa, en mai 1987, dans le cadre du 55e congrès de l'Association canadienne-française pour l'avancement des sciences. Pour une deuxième année consécutive, le colloque était parrainé par le Bureau de la statistique du Québec. La diversité des organismes ayant participé au colloque et des thèmes présentés lui assurèrent un franc succès. En effet, environ cent vingt-cinq personnes ont assisté aux trente-trois présentations.

Nous tenons à remercier tous ceux qui nous ont fait parvenir le texte de leur présentation. Nous espérons que la diffusion de ce recueil permettra aux statisticiens de profiter des expériences et recherches ici démontrées. Le Bureau de la statistique du Québec laisse aux auteurs la responsabilité entière du contenu de leur texte.

Responsables du colloque et
coordonnatrices de la publication:

Françoise Tarte
Martine Désilets
Bureau de la statistique du Québec
117 St-André
Québec (Qc)
G1K 3Y3
Tél.: (418) 643-1493

Table des matières

OUVERTURE

P.-Gendreau, Nicole, Directrice générale du Bureau de la statistique du Québec..... 1

SESSION SUR LA MÉTHODOLOGIE D'ENQUÊTES

Berthelot, Jean-Marie, Statistique Canada. Système d'interview assisté par ordinateur, un exemple..... 17

Courtemanche, Robert, et Tarte, Françoise, Bureau de la statistique du Québec. Enquête Santé-Québec: Plan de sondage..... 35

Fluet, Michel, SOM Inc. Formalisation du calcul des taux de réponse en présence d'unités hors population..... 57

Lambert, Jean, Legault, Claudine, et Maheux, Brigitte, Université de Montréal. Etude expérimentale sur l'amélioration du taux de réponse dans les enquêtes postales auprès des médecins..... 65

Landry, Rosaire, Ville de Laval. Sélection du répondant lors de sondages téléphoniques..... 77

Paquet, Pierre, Régie de l'assurance automobile du Québec. Enquête sur la conduite automobile avec les facultés affaiblies par l'alcool-1986..... 97

Proulx, Myriam, ministère du Revenu du Québec. Utilisation de l'échantillonnage avec probabilités proportionnelles à la taille (P.P.S.): exemple d'application sur un fichier administratif..... 109

Schiopu-Kratina, Ioana, et Srinath, K.P., Statistique Canada. L'utilisation de l'estimateur par le quotient pour mesurer les tendances..... 121

SESSION SUR LES SÉRIES CHRONOLOGIQUES

Cholette, Pierre A., Statistique Canada. L'étalonnage et l'interpolation des données dans les instituts de statistique..... 129

Laberge, Claude, Université Laval et INRS-Eau. L'estimation robuste des paramètres du modèle AR(1)..... 157

Lefrançois, Bernard, Statistique Canada. Application des séries chronologiques à l'étude des révisions..... 171

Quenneville, Benoît, Statistique Canada. Une application du filtrage de Kalman pour l'estimation des niveaux et des tendances dans les enquêtes répétées (référence)..... 181

SESSION SUR LES BANQUES DE DONNÉES

Bilodeau, Danielle, Bourque, Louise, et Tarte, Françoise, Bureau de la statistique du Québec. Utilisation de fichiers administratifs à des fins statistiques: essai d'analyse de la mobilité des jeunes travailleurs.. 183

Ouellette, Raymond, ministère de l'Education du Québec. Utilisation d'un fichier administratif à des fins statistiques: l'exemple du fichier des allocations familiales du Québec..... 201

Pampalon, Michel, Régie de l'assurance automobile du Québec. Exploitation de fichiers administratifs à des fins statistiques: regroupement des données sur les conducteurs (permis, infractions et accidents)..... 209

SESSION SUR L'ANALYSE DE DONNÉES

Béland, François, Lambert, Jean, et Maheux, Brigitte, Université de Montréal. La mesure des attitudes et des comportements dans les enquêtes par questionnaires..... 231

Bernier, Julie, Université Laval. Analyse graphique des tableaux de fréquences..... 243

Lambert, Jean, Université de Montréal. Quelques réflexions sur l'inférence statistique dans l'analyse de données catégorielles..... 251

Montégiani, Mario, Perreault, Luc, et Simard, Serge, Université Laval. Portrait statistique de peuplements résineux ayant subi des traitements sylvicoles..... 261

DISCOURS PRONONCÉ

PAR

MADAME NICOLE GENDREAU

DIRECTRICE GÉNÉRALE

DU

BUREAU DE LA STATISTIQUE DU QUÉBEC

À L'OUVERTURE DU

COLLOQUE SUR LES APPLICATIONS DE LA STATISTIQUE

QUI S'EST DÉROULÉ DANS LE CADRE DU

55E CONGRÈS ANNUEL

DE

L'ASSOCIATION CANADIENNE-FRANÇAISE POUR L'AVANCEMENT DES SCIENCES

UNIVERSITÉ D'OTTAWA

19 MAI 1987

Diapositive # 1

Mesdames les organisatrices,
Mesdames,
Messieurs,

Il me fait plaisir d'être avec vous cet après-midi et d'inaugurer, dans le cadre du 55e Congrès annuel de l'Association canadienne-française pour l'avancement des sciences, ce colloque sur les app' ications de la statistique.

Diapositive # 2

Ce colloque que parraine le Bureau de la statistique du Québec, je suis personnellement fière d'y participer. C'est un événement d'importance pour toute la communauté statistique francophone. Et je suis fière de voir l'organisme que je dirige, le Bureau de la statistique du Québec, s'associer pour une deuxième année consécutive, aux objectifs que, sans nul doute, vous poursuivez tous: la diffusion et la promotion d'une information statistique de qualité.

Je suis par ailleurs heureuse de constater que le milieu universitaire a répondu à l'invitation lancée l'an dernier pour faire de ce colloque le rendez-vous par excellence des statisticiens d'expression française.

Il m'apparaît grandement souhaitable que les gens des milieux de la recherche universitaire et ceux du secteur privé puissent rencontrer les statisticiens oeuvrant dans les administrations publiques municipales ou dans les administrations fédérale et québécoise. Et je veux espérer qu'avec ce Colloque sur les applications de la statistique débute une nouvelle ère de collaboration et d'échanges entre nous tous. Je vous souhaite donc à chacun que cette rencontre porte fruit et, plus particulièrement, que les échanges et la collaboration qui s'instaureront lors des trois jours de ce colloque soient bien vivants et se poursuivent toute l'année durant.

Diapositive # 3

Sans plus tarder, je profiterai maintenant de la tribune de choix qui m'est offerte dans le cadre de ce colloque sur les applications de la statistique, pour annoncer qu'en décembre 1987 et tout au long de l'année 1988, le Bureau de la statistique du Québec soulignera ses soixante-quinze ans d'existence sous le thème de: Soixante-quinze ans de fiabilité!

Soixante-quinze ans de fiabilité, parce qu'au cours de ses soixante-quinze ans d'existence, le Bureau de la statistique du Québec a démontré qu'il était une institution digne de confiance. Une institution dont les données sont fiables.

SOIXANTE-QUINZE ANS DE FIABILITÉ, PARCE QUE, GRÂCE À LA COMPÉTENCE DE SON PERSONNEL, LA QUALITÉ DE SES PRODUITS, DE SES DONNÉES, A TOUJOURS FAIT LA FORCE DU BUREAU DE LA STATISTIQUE DU QUÉBEC.

DIAPO-
SITIVE
4

POUR SOULIGNER LE SOIXANTE-QUINZIÈME ANNIVERSAIRE DU BUREAU DE LA STATISTIQUE DU QUÉBEC, JE ME PERMETTRAI DONC, CET APRÈS-MIDI, DE VOUS EN RELATER UNE TRANCHE D'HISTOIRE. JE VOUS PARLERAI TOUT PARTICULIÈREMENT DES CIRCONSTANCES QUI ONT ENTOURÉ SA CRÉATION.

MAIS AVANT, JE CROIS QUE VOUS SEREZ INTÉRESSÉS D'APPRENDRE QUE LE PREMIER RECENSEMENT DE LA POPULATION DU CANADA A ÉTÉ RÉALISÉ EN 1665-1666, SOUS L'AUTORITÉ DE L'INTENDANT TALON. EN FAIT, IL FAUDRAIT PLUTÔT PARLER DU PREMIER RECENSEMENT DE LA POPULATION DE LA NOUVELLE-FRANCE, VOUS L'AUREZ TOUS COMPRIS ...

EN NOUVELLE-FRANCE DONC, LES RECENSEMENTS SE SONT SUCCÉDÉS À INTERVALLES IRRÉGULIERS. LE SECOND RECENSEMENT A EU LIEU EN 1667, LE TROISIÈME EN 1679, LE QUATRIÈME EN 1681. LES RECENSEMENTS ET LA STATISTIQUE NE DATENT PAS D'HIER AU QUÉBEC ...

PASSONS, SANS PLUS TARDER, AUX CIRCONSTANCES ENTOURANT LA CRÉATION DU BUREAU DES STATISTIQUES DE QUÉBEC.

L'ADOPTION DE LA LOI CRÉANT LE BUREAU DES STATISTIQUES DE QUÉBEC, EN 1912, S'EST RÉALISÉE DANS UN CONTEXTE QUI N'EST PAS SANS RAPPELER CERTAINS ÉPISODES MODERNES DE LA VIE POLITIQUE. EN FAIT, LE QUÉBEC APPRÉHENDAIT CERTAINEMENT À L'ÉPOQUE, LE RÉSULTAT DES ENQUÊTES ET SÉANCES D'ÉTUDES DE LA COMMISSION FOSTER. LA COMMISSION FOSTER, DU NOM DE SON PRÉSIDENT, LE MINISTRE FÉDÉRAL DU COMMERCE, L'HONORABLE GEORGES E. FOSTER, ÉTAIT CHARGÉE D'ENQUÊTER SUR LES STATISTIQUES OFFICIELLES DU CANADA.

PLUS PRÉCISÉMENT, ELLE DEVAIT FAIRE ENQUÊTE SUR LE TRAVAIL STATISTIQUE QUI S'ACCOMPLISSAIT DANS LES DIVERS MINISTÈRES FÉDÉRAUX, SUR L'ÉTENDUE DU TRAVAIL, LES MÉTHODES EMPLOYÉES, L'EXACTITUDE OBTENUE, POUR SAVOIR SI LE MÊME TRAVAIL ÉTAIT RÉPÉTÉ INUTILEMENT ET JUSQU'À QUEL POINT IL L'ÉTAIT. DISONS TOUT DE SUITE QUE SEUL LE DÉNOMBREMENT ET LA COMPILATION DES RAPPORTS DES RECENSEMENTS DÉCENNAUX SEMBLAIENT SATISFAIRE LE MINISTRE FOSTER. CE DERNIER ÉTAIT PAR AILLEURS AUTORISÉ À CORRESPONDRE AVEC LES DIVERS GOUVERNEMENTS PROVINCIAUX EN VUE DE DÉTERMINER QUEL GENRE DE TRAVAIL STATISTIQUE ÉTAIT EXÉCUTÉ PAR LES PROVINCES, LES MÉTHODES SUIVIES, LES RÉSULTATS ATTEINTS, ET POUR ÉTABLIR JUSQU'À QUEL POINT LES PROVINCES POUVAIENT AIDER OU RÉPÉTER LE TRAVAIL EXÉCUTÉ À L'ÉPOQUE PAR LE GOUVERNEMENT FÉDÉRAL.

Selon les termes de l'arrêté en conseil, la Commission devait soumettre son rapport sur un système de statistiques générales simple, en rapport avec les besoins du Canada et rencontrant les demandes de l'époque. Le rapport fut déposé le 1er décembre 1912.

Au Québec, alors même que la commission fédérale tenait des séances d'études, le lieutenant-gouverneur annonçait, dans le discours du trône, le 5 novembre 1912, un projet de loi concernant la création d'un bureau des statistiques.

Quelques semaines plus tard, le Secrétaire de la province, l'honorable Jérémie-Louis Décarie, déposait le projet de loi établissant le Bureau des statistiques du Québec. La loi sanctionnnée le 21 décembre 1912 ne devait toutefois entrer en vigueur que le 11 novembre 1913.

Cette loi décrétait que, sous l'autorité du Secrétaire de la province, le chef du "Bureau des statistiques de Québec" pouvait recueillir, condenser et mettre en tableau des statistiques et renseignements utiles sur le Québec. Ces statistiques et renseignements utiles portaient tout particulièrement à l'époque sur l'éducation, l'industrie, le commerce, l'agriculture, la population, la colonisation et les produits naturels du sol.

J'AIMERAIS, PAR AILLEURS, VOUS MENTIONNER IMMÉDIATEMENT QUE LA CRÉATION DU BUREAU DES STATISTIQUES DE QUÉBEC, EN 1912, EN FAIT LE PLUS VIEIL ORGANISME CENTRAL STATISTIQUE AU CANADA, PUISQUE CE N'EST QU'EN 1918 QUE LE GOUVERNEMENT DU CANADA AGRÉAIT À LA PREMIÈRE ET À LA PLUS IMPORTANTE DES RECOMMANDATIONS DE LA COMMISSION FOSTER, CELLE DE CRÉER LE "BUREAU CENTRAL DES STATISTIQUES" FÉDÉRAL. MAIS REVENONS-EN TOUT DE SUITE À LA CRÉATION DU BUREAU DES STATISTIQUES DE QUÉBEC.

LA LOI QUÉBÉCOISE FUT SANCTIONNÉE EN DÉCEMBRE 1912. POUR METTRE SUR PIED LE BUREAU DE QUÉBEC, LE PREMIER MINISTRE DE LA PROVINCE, SIR LOMER GOUIN, S'EST ALORS ADRESSÉ AU MINISTRE FRANÇAIS DES AFFAIRES ÉTRANGÈRES, MONSIEUR GABRIEL HANOTAUX, AFIN D'OBTENIR LA PARTICIPATION DE LA FRANCE. ACQUIESÇANT À LA DEMANDE DU PREMIER MINISTRE DU QUÉBEC ET DU MINISTRE FRANÇAIS DES AFFAIRES ÉTRANGÈRES, LE DIRECTEUR DE LA STATISTIQUE DE FRANCE, LUCIEN MARCH, CONSENTIT À CE QU'UN DE SES ADJOINTS VIENNE AU QUÉBEC POUR Y FONDER UN BUREAU DES STATISTIQUES. CET ADJOINT QUI VINT EN MISSION AU QUÉBEC CE FUT HENRI RUNLE, UN LICENCIÉ ES-SCIENCES DE LA SORBONNE EN 1907, QUI ÉTAIT ARRIVÉ PREMIER AU CONCOURS DE STATISTICIENS À LA "STATISTIQUE GÉNÉRALE DE LA FRANCE" EN NOVEMBRE 1907.

DIAPOSITIVE # 5

Diapositive # 6

À l'oeuvre dès septembre 1913, soit avant même que la nouvelle loi établissant le Bureau des statistiques du Québec ne soit en vigueur, Henri Bunle put remettre au Secrétaire de la province, son Rapport sur l'organisation d'un bureau de la statistique après à peine neuf mois de travail, soit le 27 juin 1914.

Le rapport de Henri Bunle, dont le Bureau de la statistique du Québec conserve l'unique exemplaire, montre la nécessité, l'importance et l'utilité d'un service de statistiques. Il traite de l'organisation d'un bureau et du mode de recueillement des statistiques. Il suggère des travaux à faire pour établir des statistiques nouvelles et pour améliorer les statistiques déjà existantes. Le rapport Bunle énumère même les différentes publications que le Bureau des statistiques se devait, selon l'auteur, de faire paraître dès les premières années. Au nombre de ces publications, figure un Annuaire statistique.

Diapositive # 7

Henri Bunle attachait une importance toute particulière à la publication d'un annuaire statistique. Parallèlement à la rédaction de son rapport, il préparait d'ailleurs le premier Annuaire statistique de Québec dont le plan a servi de base à de nombreux annuaires québécois par la suite.

Henri Bunle considérait qu'un annuaire statistique devait présenter sous forme claire et condensée, le résumé des statistiques recueillies par les services publics des divers ministères provinciaux et fédéraux. Utile aux gouvernants comme aux gens d'étude, cette publication réunissant en un seul volume toutes les statistiques d'ordre général permettait, selon Bunle, d'avoir un aperçu exact et complet de la physionomie du Québec, de ses ressources et de son développement historique. Elle devait également mieux faire connaître le Québec à l'extérieur. Parlant de la publication d'un annuaire, Henri Bunle disait justement: "Il y a certainement là un très intéressant, très utile et très puissant moyen de propagande, que ne doit pas négliger un pays jeune, aux nombreuses ressources, encore inexploitées malheureusement. Ce moyen d'action très énergique doit attirer à la province de Québec la main-d'oeuvre et les capitaux étrangers: si l'on veut être connu, il faut se faire connaître, rappeler constamment que l'on vit, indiquer ce que l'on fait".

Henri Bunle avait raison de penser que le type d'annuaire qu'il proposait permettrait à ses utilisateurs d'avoir un portrait fidèle et complet du Québec. En fait, aujourd'hui encore, les historiens recueillent dans les premiers annuaires des mines de renseignements auxquels ils

N'AURAIENT PEUT-ÊTRE JAMAIS EU ACCÈS AUTREMENT. ILS RECUEILLENT DES DONNÉES FACTUELLES, MAIS CE QUI EST PROBABLEMENT PLUS IMPORTANT POUR EUX, ILS Y PRENNENT LE POULS DE L'ÉPOQUE.

...

VOUS COMPRENDREZ FACILEMENT SI JE VOUS DIS QUE DANS L'ANNUAIRE DE 1914, ON TRAITE DE LA COLONISATION ET NON PAS DE L'AGRICULTURE; DE L'INSTRUCTION PLUTÔT QUE DE L'ÉDUCATION. IL Y EST AUSSI QUESTION D'UNE LOI QUI ACCORDE UNE "ASSISTANCE AUX FAMILLES NOMBREUSES", C'EST-À-DIRE AUX FAMILLES AYANT DOUZE ENFANTS BEL ET BIEN VIVANTS. TOUT UNE AUTRE ÉPOQUE! N'EST-CE PAS? MAIS REVENONS-EN AU FONDATEUR DU BUREAU DES STATISTIQUES DE QUÉBEC.

HENRI BUNLE NE PUT ASSISTER À LA SORTIE DE LA PREMIÈRE ÉDITION DE L'ANNUAIRE STATISTIQUE DU QUÉBEC EN OCTOBRE 1914. EN FAIT, IL VENAIT TOUT JUSTE DE REGAGNER LA FRANCE, EN JUILLET 1914, DÉSIREUX D'Y PASSER SES VACANCES. CELLES-CI FURENT SÛREMENT ÉCOURTÉES CEPENDANT, PUISQUE LE 2 AOÛT 1914, HENRI BUNLE ÉTAIT APPELÉ SOUS LES DRAPEAUX À LA SUITE DE LA DÉCLARATION DE LA GUERRE.

DANS UNE ALLOCUTION QU'IL PRONONÇAIT EN MARS 1966, HENRI BUNLE DISAIT QU'AVANT LA PREMIÈRE GUERRE MONDIALE "LA STATISTIQUE ÉTAIT UNE PARENTE PAUVRE ET DÉCONSIDÉRÉE", QU'ELLE

AVAIT MAUVAISE PRESSE ET QU'UN "MINISTRE FRANÇAIS NE CRAIGNAIT PAS D'AFFIRMER QU'ELLE ÉTAIT NON SEULEMENT INUTILE MAIS NUISIBLE". L'ÉPOQUE ÉTAIT BIEN DIFFICILE, ON PEUT LE CONSTATER, POUR LA STATISTIQUE ET LES STATISTICIENS.

AU QUÉBEC, HENRI BUNLE NE REÇUT D'AILLEURS PAS DE TRAITEMENT DE FAVEUR EN 1913 ET 1914. IL N'AVAIT NI BUREAU, NI DACTYLOGRAPHE, NI COLLABORATEURS, SAUF DANS LES TOUT DERNIERS MOIS, POUR L'ANNUAIRE STATISTIQUE DE QUÉBEC. IL TRAVAILLAIT TOUT SIMPLEMENT DANS UN HÔTEL DE QUÉBEC. C'EST CE QU'IL NOUS A LUI-MÊME MENTIONNÉ DANS UNE INTERVIEW-VIDÉO RÉALISÉE ALORS QU'IL ÉTAIT ÂGÉ DE 102 ANS, SOIT À PEINE TROIS MOIS AVANT SON DÉCÈS SURVENU EN MARS 1986.

HUMBLES DÉBUTS, DONC, POUR LA STATISTIQUE AU QUÉBEC, MAIS AVEC LES ANNÉES, LE CHAMP D'ACTIVITÉS DU BUREAU S'EST ÉLARGI ET LE NOMBRE DE PUBLICATIONS A CONSIDÉRABLEMENT AUGMENTÉ.

LE BUREAU DE LA STATISTIQUE DU QUÉBEC D'AUJOURD'HUI CONTINUE TOUTEFOIS DE COLLABORER AVEC CE QUE HENRI BUNLE APPELAIT LES "SERVICES PUBLICS DES DIVERS MINISTÈRES PROVINCIAUX ET FÉDÉRAUX". SON INTERLOCUTEUR FÉDÉRAL EST STATISTIQUE CANADA ET LA COORDINATION DU SYSTÈME STATISTIQUE QUÉBÉCOIS EST UN DES RÔLES LES PLUS FONDAMENTAUX QUE JOUE LE BUREAU DE LA STATISTIQUE DU QUÉBEC.

UNE COMPLÉMENTARITÉ S'EST D'AILLEURS ÉTABLIE AU FIL DES ANS ENTRE LES OPÉRATIONS DE COLLECTE DES DONNÉES DES DIVERS INTERVENANTS DU SYSTÈME STATISTIQUE QUÉBÉCOIS, À SAVOIR LE BUREAU DE LA STATISTIQUE DU QUÉBEC, STATISTIQUE CANADA ET LES MINISTÈRES ET ORGANISMES DU GOUVERNEMENT DU QUÉBEC. DE NOMBREUSES ENTENTES D'ÉCHANGE DE DONNÉES, FAITES EN VERTU DES LOIS, PERMETTENT MAINTENANT DE LIMITER LA DUPLICATION DANS LA CUEILLETTE DE RENSEIGNEMENTS STATISTIQUES.

DIAPOSITIVE # 8

EN COMPARAISON AVEC STATISTIQUE CANADA, LE BUREAU DE LA STATISTIQUE DU QUÉBEC FOURNIT UNE INFORMATION PLUS DÉTAILLÉE ET MIEUX INTÉGRÉE SUR LE QUÉBEC, SES RÉGIONS, SES SECTEURS D'ACTIVITÉ ÉCONOMIQUE, SA POPULATION ET SES GROUPES SOCIO-ÉCONOMIQUES. CELA LUI PERMET ENTRE AUTRES DE RÉPONDRE AUX BESOINS DES ADMINISTRATIONS PUBLIQUES ET PARA-PUBLIQUES QUI DOIVENT DISPOSER D'UNE CONNAISSANCE PLUS SPÉCIFIQUE SUR LA RÉALITÉ QUÉBÉCOISE ET SES COMPOSANTES.

DIAPOSITIVE # 9

LE BUREAU DE LA STATISTIQUE DU QUÉBEC OFFRE, EN PLUS DES OPÉRATIONS DE COLLECTE, D'ANALYSE ET D'INTÉGRATION DES DONNÉES, LA COMPÉTENCE DE SES EXPERTS EN MÉTHODOLOGIE D'ENQUÊTE, EN MÉTHODOLOGIE QUANTITATIVE, EN NORMALISATION ET EN CLASSIFICATION. CES SERVICES SONT PRINCIPALEMENT OFFERTS AUX DIVERS MINISTÈRES ET ORGANISMES DU GOUVERNEMENT DU QUÉBEC.

Si le Bureau de la statistique du Québec s'efforce de répondre, en tout premier lieu, aux besoins des administrations publiques et para-publiques, n'en concluez surtout pas cependant qu'il ignore les besoins des entreprises et de la population en général.

Diapositive # 10

Il leur offre notamment de réaliser des études d'impact grâce aux modèles d'analyse qu'il a développés. Et il met à la disposition de la population en général aussi bien que des entreprises ou de l'Administration un centre d'information et de documentation, à Québec et à Montréal.

Diapositive # 11

Voilà pour le champ d'action du Bureau de la statistique du Québec et pour ses diverses clientèles. Passons maintenant à ses différents secteurs d'activités.

Diapositive # 12

Les domaines d'intervention du Bureau de la statistique du Québec sont les domaines économique, social, démographique, culturel et environnemental.

L'état d'avancement des travaux est toutefois bien inégal d'un domaine à l'autre. Cela se comprend facilement si l'on songe que tout au long de son histoire, le Bureau de la statistique du Québec a toujours tenu à répondre aux besoins et aux préoccupations de son époque. Et cela se

comprend d'autant plus lorsque l'on analyse les rattachements administratifs du Bureau de la statistique du Québec.

Sous l'autorité du Secrétaire de la province dès 1912, le Bureau fut rattaché en 1930 au ministère des Affaires municipales que dirigeait alors l'honorable Louis-Alexandre Taschereau, également premier ministre du Québec. Le Bureau resta partie intégrante du ministère des Affaires municipales jusqu'en juillet 1936. Cependant, il faut préciser qu'en mai 1935, le ministère des Affaires municipales avait fait corps avec l'Industrie et le Commerce. D'ailleurs, en 1936, le Bureau fut rattaché, quelque deux mois, au nouveau ministère du Commerce et de l'Industrie avant que celui-ci ne fasse à nouveau corps avec les Affaires municipales pour sept ans. De mars 1943 à septembre 1981, le Bureau fut à nouveau placé sous l'autorité du ministère de l'Industrie et du Commerce. En septembre 1981, le Bureau se greffait au ministère du Conseil exécutif et depuis mars 1983, il relève du ministère des Finances.

Diapositive # 13

En terminant, je vous invite à visionner jeudi, le 22 mai à 17 heures, l'interview-vidéo de Henri Bunle qu'a réalisée l'université de Montréal en collaboration avec le Bureau de la statistique du Québec. Lors de l'historique interview, Henri Bunle était âgé de 102 ans et son passage au Québec remontait à quelque soixante-treize ans. C'est un hommage que nous voulons rendre au fondateur du Bureau en ce soixante-quinzième anniversaire.

Je veux, par ailleurs, remercier les organisatrices de ce Colloque, mesdames Françoise Tarte et Martine Désilets, du Bureau de la statistique du Québec. Elles n'ont pas compté les heures, ni ménagé leurs efforts pour voir à la bonne marche de ce colloque.

À tous, enfin, je souhaite de fructueux échanges au cours de ce colloque en espérant, bien sûr, qu'ils se poursuivent toute l'année durant.

NOTE: La rédaction de ce texte s'inspire des données de l'Annuaire statistique du Québec de 1914 et des travaux de recherches et écrits d'un ancien directeur du Bureau de la statistique du Québec, J.-C. McGee (Annuaire statistique du Québec, 1953), et de l'actuel secrétaire du Bureau, monsieur Lionel LaBerge (Annuaire statistique du Québec, 1970).

SYSTÈME D'INTERVIEW ASSISTÉ PAR ORDINATEUR, UN EXEMPLE.[1]

Jean-Marie Berthelot[2]

Résumé

L'utilisation de l'ordinateur pour faciliter et améliorer la collecte des données est assez récente dans le domaine des enquêtes économiques. Le présent article décrit, de façon générale, le système d'interview téléphonique assisté par ordinateur qui a été développé pour l'enquête mensuelle sur les stocks, livraisons et commandes des industries manufacturières.

1. HISTORIQUE

L'Enquête mensuelle sur les Stocks, Livraisons et Commandes des industries manufacturières est une enquête échantillon couvrant environ 11,000 établissements manufacturiers au Canada. Cette enquête existe depuis la fin des années 1960. Dès 1973, des efforts ont été faits en vue de la régionalisation de la collecte des données. Ceux-ci se sont traduits, en 1978, par l'implantation d'un système de collecte des données dans tous les bureaux régionaux. En 1981, un système de saisie des données, accompagné d'un système de transmission électronique des données au siège social, a été développé et son implantation complétée dans les 8 bureaux régionaux de Statistique Canada en 1983. L'implantation de ce système a permis des économies significatives, en éliminant le besoin de photocopier et de transcrire manuellement plusieurs documents.

Suite au succès de la régionalisation de 1983, un projet s'amorçait en janvier 1984, ayant comme objectifs:

i) de considérer le développement d'un système automatique de gestion et de contrôle des documents pour ESLC pour utilisation dans les bureaux régionaux,

ii) de déterminer les fonctions additionnelles de l'ESLC qui pourraient être effectuées par les bureaux régionaux, et

iii) de considérer la faisabilité de l'utilisation d'un système d'interview interactif pour l'enquête.

---

[1] *Présenté au colloque sur les applications statistiques du 55e congrès de l'Association Canadienne Française pour l'Avancement des Sciences, mai 87.*

[2] *Jean-Marie Berthelot, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa(Ontario), K1A 0T6.*

Une équipe multidisciplinaire, comprenant entre autre, une firme de consultants informatique, a entrepris le travail nécessaire à l'analyse de ces objectifs. Le rapport de l'équipe de projet, présenté en mai 1984 au comité de gestion des opérations, recommandait:

i) de combiner les fonctions d'interview et de saisie des données,

ii) de régionaliser la vérification des données brutes ainsi que le maintien de l'information administrative de l'échantillon, et

iii) de produire des rapports de gestion et de contrôle, afin de faciliter la gestion de l'enquête.

Le Comité a approuvé les recommandations et a permis au projet de régionalisation d'aller de l'avant.

2. Description des fonctions du système.

Pour l'ESLC, la collecte de l'information est effectuée soit par téléphone (environ 70%), soit par la poste, suivant la préférence des répondants. Que l'information soit obtenue par téléphone ou par la poste, elle est saisie directement par l'intervieweur, en utilisant le système d'Interview Assisté par Ordinateur(IAO). Le système IAO peut être séparé en 5 fonctions distinctes: une fonction de planification automatique des entrevues, une fonction de mise à jour de l'information administrative de l'échantillon, une fonction de saisie et de vérification interactive des données, une fonction produisant des rapports pour la gestion et une fonction de transmission électronique des données.

A. Planification automatique des entrevues.

Lors de l'installation du système IAO, tous les répondants ont été contactés, de façon à connaitre, à l'intérieur d'un mois, la journée et l'heure qui leur conviendraient le mieux pour rapporter l'information receuillie par l'Enquête sur les Livraisons, Stocks et Commandes des industries manufacturières. Cette information est utilisée par le système IAO pour faire la composition automatique des numéros de téléphone, et ainsi planifier la tâche des intervieweurs.

Le système a deux modes de fonctionnement. Le mode manuel, qui permet de contacter une entreprise spécifique, peu importe l'heure ou la journée, et le mode automatique, qui utilise les possibilités de planification automatique. Le mode manuel est aussi utilisé pour saisir l'information des unités déclarantes qui préfèrent rapporter par la poste.

En mode automatique, le système gère les numéros occupés et les absents de façon à pouvoir planifier les rappels. De plus, il crée une file d'attente lorsque le temps pour compléter une entrevue est plus long qu'initialement prévu.

B. Mise à jour de l'information administrative.

A chaque mois, lorsqu'un intervieweur rejoint un répondant, l'information administrative de l'établissement (i.e. le nom, l'adresse, et la personne contact) est vérifiée. Si des changements ont eu lieu, l'intervieweur enregistre les modifications qui sont alors utilisées par le système pour mettre à jour automatiquement l'unité déclarante appropriée.

C. Saisie et vérification interactive des données.

Après avoir vérifié, et si nécessaire corrigé, l'information administrative de l'unité déclarante, l'intervieweur débute alors la partie principale de l'entrevue, consistant à recueillir et à saisir les données statistiques. Le système IAO a été conçu pour vérifier interactivement l'information recueillie par l'intervieweur. Le but de la vérification interactive des données brutes est d'améliorer la qualité des données, de réduire le fardeau de réponse en minimisant le nombre de rappels, ainsi que de minimiser le temps consacré à la collecte et à la saisie des données.

Le système de vérification des données implanté dans les bureaux régionaux effectue cinq types de vérification:

i) une vérification de cohérence

Pour l'ESLC, les variables qui doivent être recueillies pour une unité déclarante sont connues par l'entremise de la base de sondage (Recensement des Manufacturiers). La vérification de cohérence s'assure qu'aucune de ces variables n'est manquante lors de la saisie des données et que les caractéristiques qui sont rapportées sont celles qui devraient l'être. Si, en raison de certains changements dans l'entreprise, une variable ne peut plus être rapportée, (ou une nouvelle variable peut l'être), l'intervieweur peut mettre à jour la liste des variables à recueillir. On vérifie aussi la validité des zéros et on effectue le contrôle des rapports saisonniers.

ii) concordance des totaux

Si une caractéristique est séparée en sous-caractéristiques, on s'assure que la somme des sous-caractéristiques est égale à la

valeur de la caractéristique. Exemple, la somme de tous les types de livraisons est égale à la valeur des livraisons totales.

iii) Identification des données aberrantes

L'objectif de cette vérification est d'identifier les unités pour lesquelles les valeurs rapportées pour le mois courant sont anormalement élevées ou anormalement faibles, comparativement aux données rapportées pour le mois précédent. La notion de Changement Relatif Ajusté (CRA) est utilisée pour cette vérification. Le CRA est une fonction du changement mensuel en pourcentage d'une caractéristique et de la valeur de la caractéristique. La notion de CRA est utilisée, car on a constaté que la variabilité du changement mensuel était fonction de la valeur de la caractéristique. La fonction CRA permet de transformer les données de façon à réduire la dépendance de la variabilité par rapport à la caractéristique étudiée.

La fonction CRA utilisée pour l'ESLC est:

$$CRA = Changement * [Max(Courant,Précédent)/100]^{0.25}$$

où Courant= Valeur courante de la caractéristique,
Précédent= Valeur précédente de la caratéristique, et
Changement= 100 * (Courant-Précédent)/Min(Courant,Précédent)

Cette transformation représente un équilibre entre l'approche intuitive des experts de la Division de l'Industrie et la réduction de la dépendance de la variabilité de la caractéristique étudiée.

Une transaction est considérée aberrante si la valeur du CRA est à l'extérieur de certaines bornes. Les bornes sont spécifiques à chaque bureau régional et à chaque période de collecte . Les bornes sont calculées deux fois l'an en utilisant l'information recueillie au cours de la dernière année de l'enquête. Les bornes sont identifiées en utilisant la méthode standard des quartiles:

Borne inférieure= Médiane(CRA) - k*D1(CRA)

Borne supérieure= Médiane(CRA) + k*D3(CRA)

où D1= Médiane(CRA) - 1er_quartile(CRA)
D3= 3e_quartile(CRA) - Médiane(CRA), et
K est une constante permettant de contrôler l'étendue des bornes.

iv) Nouvelles commandes reçues

A chaque mois, les commandes en carnet sont augmentées par les nouvelles commandes reçues et diminuées par la valeur des livraisons. Cette vérification s'assure que la valeur des commandes en carnet de la période précédente est inférieure à la somme des livraisons et des commandes en carnet pour le mois courant, assurant ainsi que les nouvelles commandes reçues sont positives.

v) Rapports cumulatifs des livraisons

Certaines entreprises rapportent les livraisons de façon cumulative au cours d'une année. Pour s'assurer que de tels cas sont corrigés, on vérifie le nombre successif d'augmentations dans les livraisons.

D. Rapports pour la gestion.

De façon à assurer une bonne gestion et un bon contrôle de l'enquête, les huit rapports suivants sont produits:

- une **liste maîtresse de l'échantillon**, contenant la liste de tous les établissements qui doivent être contactés par un bureau régional,

- une **liste de mise à la poste**, contenant la liste des établissements devant rapporter par la poste. Cette liste est aussi utilisée pour contrôler la mise à la poste et la réception des questionnaires,

- une **liste des naissances**, identifiant les nouveaux établissements qui doivent être contactés pour un mois donné,

- un **rapport sur la planification des entrevues**, permettant au gestionnaire de connaître la liste des établissements qui doivent être contactés par téléphone, pour chaque jour de la période de collecte . Ce rapport est produit pour chacune des périodes de collecte ,

- un **rapport statistique sur les entrevues**, permettant de connaître le nombre d'entrevues complétées par un intervieweur, l'état de chacune des unités déclarantes, ainsi que le temps moyen pour compléter la collecte d'un questionnaire. Cette information permet aussi de mesurer le fardeau de travail associé à la tâche d'entrevue,

- un **sommaire de la période de collecte** , fournissant de l'information à date sur l'état de chacune des unités déclarantes. Ce rapport peut être produit à n'importe quel moment de la période de collecte,

- un **rapport détaillé de la période de collecte** , identifiant chacun des répondants, en fonction de son état de réponse (donnée valide, en erreur ou manquante). L'objectif est de permettre au gestionnaire d'identifier des patrons de réponse d'une période de référence à une autre, et

- un **rapport mensuel des activités**, fournissant l'information nécessaire au gestionnaire d'un bureau régional pour comparer les résultats de la collecte des données de son bureau régional avec les résultats des autres bureaux régionaux.

E. Transmission électronique des données.

Tout au long de la période de collecte , les données sont transmises de façon régulière au siège social. Le logiciel déjà existant dans les bureaux régionaux est utilisé à cette fin.

3. Implantation

Le système a été développé avec les mini ordinateurs Honeywell, en utilisant les logiciels: «*Data Entry Facility (DEF)*», «*COBOL avancé*», «*SORT/MERGE*» et le *langage de contrôle Honeywell*. Le système a été conçu pour être facile d'opération, par l'utilisation de menus.

Une version initiale, mais complète du système IAO, a été implanté au bureau régional de Halifax en mars 1987, de façon à effectuer un essai complet du système avant son implantation dans les autres bureaux régionaux. Les principaux objectifs de cet essai étaient:

i) de vérifier le fonctionnement du système avec des données réelles,

ii) de vérifier la qualité des données résultant de l'utilisation du nouveau système,

iii) de vérifier les outils de gestion et de contrôle du système,

iv) d'obtenir la réaction des intervieweurs concernant le sytème IAO, et

v) d'identifier les points qui pourraient être améliorés avant l'implantation dans les autres bureaux régionaux.

Le test a duré 5 mois et a permis d'identifier certains problèmes et certaines lacunes du système. L'application du test a permis de modifier le système, avant son implantation dans les autres bureaux régionaux. Les principales améliorations apportées au système ont été:

i) d'améliorer le système de re-planification des entrevues, lors d'un «non-contact»,

ii) de rendre plus facile, aux intervieweurs, l'utilisation du système, le rendre plus humain,

iii) d'améliorer le traitement des non-réponses et des rapports saisonniers,

iv) de rendre plus clair et plus fonctionnel l'identification des variables qui doivent être rapportées par un répondant,

v) de permettre l'impression, dans les bureaux régionaux, des étiquettes pour le guide des répondants,

vi) de faciliter la correction des erreurs faites sur des rapports précédents, et

vii) après une certaine date, de permettre la planification automatique d'entrevues téléphoniques pour les unités déclarant par la poste.

4. Qualité des données

Afin de vérifier la qualité des données produites par le système IAO, une collecte parallèle des données a été faite avec l'ancien système. La comparaison des données de l' ancien système avec celles du nouveau a montré que ce dernier ne cause aucune détérioration des données. Les observateurs du bureau régional de Halifax croient que le nouveau système a permis d'améliorer la qualité des données, en assurant une vérification plus structurée et systématique des données.

Une étude supplémentaire, comparant le taux et la consistance des vérifications du nouveau et de l'ancien système pour les bureaux régionaux de Halifax et de Winnipeg a été réalisée.

i) Pour l'ancien système

Les intervieweurs devaient, de façon manuelle, identifier et vérifier toutes les transactions montrant un changement mensuel supérieur à 25 %. Toutes les transactions qui ont été vérifiées par un intervieweur , entre les mois d'octobre 1985 et d'avril 1986 ont été envoyées au siège social.

ii) Pour le nouveau système

Le système de détection des données aberrantes qui devait être implanté dans les bureaux régionaux a été appliqué aux données initiales pour la même période de référence.

Le taux de rejet, ainsi que la variation de ce taux pour différents mois ont été calculés pour les deux systèmes. De plus, pour l'ancien système, le taux espéré de rejet pour l'ensemble de la période a aussi été calculé. Des résultats partiels pour la caractéristique Livraisons sont présentés dans la table suivante:

TAUX DE REJET OBSERVÉ

MOIS DE RÉFÉRENCE	HALIFAX		WINNIPEG	
	ANCIEN	NOUVEAU	ANCIEN	NOUVEAU
Oct. 85	26.4	16.2	31.8	11.2
Nov. 85	32.7	12.7	37.4	12.7
Dec. 85	33.7	18.3	42.4	13.4
Jan. 86	41.0	16.1	47.9	11.0
Fev. 86	--	--	45.0	16.6
Mar. 86	--	--	42.8	19.1
Avr. 86	--	--	46.2	14.2
TAUX MOYEN	33.2	15.8	41.9	14.1
TAUX ESPÉRÉ	52.0	--	52.0	--

NOTE: Voir C.S.I.O. REGIONALIZATION - PRELIMINARY REPORT, pour les résultats détaillés de cette étude.

En analysant ces résultats, on remarque que le taux de rejet pour l'ancien système varie beaucoup plus d'un mois à l'autre que celui du nouveau système. Le nouveau système semble s'adapter beaucoup mieux à la saisonnalité des données que la règle arbitraire qui consistait à rejeter les unités ayant un changement supérieur a 25 % en valeur absolue. On note aussi que le taux espéré de rejet pour l'ancien système est supérieur au taux de rejet observé par les intervieweurs. Cette situation peut avoir un effet significatif sur la qualité des données produites par l'ancien système. En effet, certaines unités qui n'ont pas été vérifiées par les intervieweurs, mais qui auraient du l'être, peuvent être en erreur.

On note aussi que le nouveau système de détection des données aberrantes identifie beaucoup moins d'unités. De façon à s'assurer que le nouveau système identifie bien les erreurs, une comparaison entre les unités rejetées par ce système et les unités qui ont été modifiées suite à une vérification des intervieweurs sous l'ancien système a été effectuée pour la variable Livraisons. Sur les 131 corrections apportées par les intervieweurs, 129 ont été identifiées par le nouveau système. On remarque donc que le nouveau système, tout en rejetant moins d'unités pour la vérification, identifie pour ainsi dire, toutes les transactions corrigées par les intervieweurs. De plus, le nouveau système étant appliqué de façon automatique et uniforme, assure un meilleur contrôle sur la qualité globale des données.

5. Conclusion

Le système d'Interview Assisté par Ordinateur pour l'Enquête sur les Stocks, Livraisons et Commandes des industries manufacturières représente une amélioration certaine sur l'ancien système de collecte et de saisie des données. Le nouveau système utilise avantageusement l'ordinateur pour améliorer la qualité des données, réduire le fardeau de réponse et diminuer les coûts associés à la collecte de l'information.

Le module de planification automatique des entrevues permet de s'adapter à la disponibilité des répondants. Le système d'interview interactif permet de réduire le nombre de rappels. Le système de détection des données aberrantes utilisé assure une meilleure qualité globale de l'information, effectuant un contrôle structuré, automatique et uniforme. Les rapports pour la gestion permettent un meilleur contrôle des opérations.

6. Bibliographie

CSIO - Regionalization working group, *CSIO regionalization preliminary report*, rapport interne, Statistique Canada, décembre 1986.

Tambay J.-L., *Study of outliers in the CSIO regional offices*, rapport interne, Statistique Canada, mars 1986.

Tambay J.-L., *Application of tolerance edits to regional office corrections*, rapport interne, Statistique Canada, juin 1986.

SYSTÈME D'INTERVIEW ASSISTÉ PAR ORDINATEUR, UN EXEMPLE.[1]

Jean-Marie Berthelot[2]

Résumé

L'utilisation de l'ordinateur pour faciliter et améliorer la collecte des données est assez récente dans le domaine des enquêtes économiques. Le présent article décrit, de façon générale, le système d'interview téléphonique assisté par ordinateur qui a été développé pour l'enquête mensuelle sur les stocks, livraisons et commandes des industries manufacturières.

1. HISTORIQUE

L'Enquête mensuelle sur les Stocks, Livraisons et Commandes des industries manufacturières est une enquête échantillon couvrant environ 11,000 établissements manufacturiers au Canada. Cette enquête existe depuis la fin des années 1960. Dès 1973, des efforts ont été faits en vue de la régionalisation de la collecte des données. Ceux-ci se sont traduits, en 1978, par l'implantation d'un système de collecte des données dans tous les bureaux régionaux. En 1981, un système de saisie des données, accompagné d'un système de transmission électronique des données au siège social, a été développé et son implantation complétée dans les 8 bureaux régionaux de Statistique Canada en 1983. L'implantation de ce système a permis des économies significatives, en éliminant le besoin de photocopier et de transcrire manuellement plusieurs documents.

Suite au succès de la régionalisation de 1983, un projet s'amorçait en janvier 1984, ayant comme objectifs:

i) de considérer le développement d'un système automatique de gestion et de contrôle des documents pour ESLC pour utilisation dans les bureaux régionaux,

ii) de déterminer les fonctions additionnelles de l'ESLC qui pourraient être effectuées par les bureaux régionaux, et

iii) de considérer la faisabilité de l'utilisation d'un système d'interview interactif pour l'enquête.

---

[1] *Présenté au colloque sur les applications statistiques du 55e congrès de l'Association Canadienne Française pour l'Avancement des Sciences, mai 87.*

[2] *Jean-Marie Berthelot, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa(Ontario), K1A 0T6.*

Une équipe multidisciplinaire, comprenant entre autre, une firme de consultants informatique, a entrepris le travail nécessaire à l'analyse de ces objectifs. Le rapport de l'équipe de projet, présenté en mai 1984 au comité de gestion des opérations, recommandait:

i) de combiner les fonctions d'interview et de saisie des données,

ii) de régionaliser la vérification des données brutes ainsi que le maintien de l'information administrative de l'échantillon, et

iii) de produire des rapports de gestion et de contrôle, afin de faciliter la gestion de l'enquête.

Le Comité a approuvé les recommandations et a permis au projet de régionalisation d'aller de l'avant.

2. Description des fonctions du système.

Pour l'ESLC, la collecte de l'information est effectuée soit par téléphone (environ 70%), soit par la poste, suivant la préférence des répondants. Que l'information soit obtenue par téléphone ou par la poste, elle est saisie directement par l'intervieweur, en utilisant le système d'Interview Assisté par Ordinateur(IAO). Le système IAO peut être séparé en 5 fonctions distinctes: une fonction de planification automatique des entrevues, une fonction de mise à jour de l'information administrative de l'échantillon, une fonction de saisie et de vérification interactive des données, une fonction produisant des rapports pour la gestion et une fonction de transmission électronique des données.

A. Planification automatique des entrevues.

Lors de l'installation du système IAO, tous les répondants ont été contactés, de façon à connaitre, à l'intérieur d'un mois, la journée et l'heure qui leur conviendraient le mieux pour rapporter l'information receuillie par l'Enquête sur les Livraisons, Stocks et Commandes des industries manufacturières. Cette information est utilisée par le système IAO pour faire la composition automatique des numéros de téléphone, et ainsi planifier la tâche des intervieweurs.

Le système a deux modes de fonctionnement. Le mode manuel, qui permet de contacter une entreprise spécifique, peu importe l'heure ou la journée, et le mode automatique, qui utilise les possibilités de planification automatique. Le mode manuel est aussi utilisé pour saisir l'information des unités déclarantes qui préfèrent rapporter par la poste.

En mode automatique, le système gère les numéros occupés et les absents de façon à pouvoir planifier les rappels. De plus, il crée une file d'attente lorsque le temps pour compléter une entrevue est plus long qu'initialement prévu.

B. Mise à jour de l'information administrative.

A chaque mois, lorsqu'un intervieweur rejoint un répondant, l'information administrative de l'établissement (i.e. le nom, l'adresse, et la personne contact) est vérifiée. Si des changements ont eu lieu, l'intervieweur enregistre les modifications qui sont alors utilisées par le système pour mettre à jour automatiquement l'unité déclarante appropriée.

C. Saisie et vérification interactive des données.

Après avoir vérifié, et si nécessaire corrigé, l'information administrative de l'unité déclarante, l'intervieweur débute alors la partie principale de l'entrevue, consistant à recueillir et à saisir les données statistiques. Le système IAO a été conçu pour vérifier interactivement l'information recueillie par l'intervieweur. Le but de la vérification interactive des données brutes est d'améliorer la qualité des données, de réduire le fardeau de réponse en minimisant le nombre de rappels, ainsi que de minimiser le temps consacré à la collecte et à la saisie des données.

Le système de vérification des données implanté dans les bureaux régionaux effectue cinq types de vérification:

i) une vérification de cohérence

Pour l'ESLC, les variables qui doivent être recueillies pour une unité déclarante sont connues par l'entremise de la base de sondage (Recensement des Manufacturiers). La vérification de cohérence s'assure qu'aucune de ces variables n'est manquante lors de la saisie des données et que les caractéristiques qui sont rapportées sont celles qui devraient l'être. Si, en raison de certains changements dans l'entreprise, une variable ne peut plus être rapportée, (ou une nouvelle variable peut l'être), l'intervieweur peut mettre à jour la liste des variables à recueillir. On vérifie aussi la validité des zéros et on effectue le contrôle des rapports saisonniers.

ii) concordance des totaux

Si une caractéristique est séparée en sous-caractéristiques, on s'assure que la somme des sous-caractéristiques est égale à la

valeur de la caractéristique. Exemple, la somme de tous les types de livraisons est égale à la valeur des livraisons totales.

iii) Identification des données aberrantes

L'objectif de cette vérification est d'identifier les unités pour lesquelles les valeurs rapportées pour le mois courant sont anormalement élevées ou anormalement faibles, comparativement aux données rapportées pour le mois précédent. La notion de Changement Relatif Ajusté (CRA) est utilisée pour cette vérification. Le CRA est une fonction du changement mensuel en pourcentage d'une caractéristique et de la valeur de la caractéristique. La notion de CRA est utilisée, car on a constaté que la variabilité du changement mensuel était fonction de la valeur de la caractéristique. La fonction CRA permet de transformer les données de façon à réduire la dépendance de la variabilité par rapport à la caractéristique étudiée.

La fonction CRA utilisée pour l'ESLC est:

$$CRA = Changement * [Max(Courant, Précédent)/100]^{0.25}$$

où Courant= Valeur courante de la caractéristique,
Précédent= Valeur précédente de la caratéristique, et
Changement= 100 * (Courant-Précédent)/Min(Courant,Précédent)

Cette transformation représente un équilibre entre l'approche intuitive des experts de la Division de l'Industrie et la réduction de la dépendance de la variabilité de la caractéristique étudiée.

Une transaction est considérée aberrante si la valeur du CRA est à l'extérieur de certaines bornes. Les bornes sont spécifiques à chaque bureau régional et à chaque période de collecte . Les bornes sont calculées deux fois l'an en utilisant l'information recueillie au cours de la dernière année de l'enquête. Les bornes sont identifiées en utilisant la méthode standard des quartiles:

Borne inférieure= Médiane(CRA) - k*D1(CRA)

Borne supérieure= Médiane(CRA) + k*D3(CRA)

où D1= Médiane(CRA) - 1er_quartile(CRA)
D3= 3e_quartile(CRA) - Médiane(CRA), et
K est une constante permettant de contrôler l'étendue des bornes.

iv) Nouvelles commandes reçues

A chaque mois, les commandes en carnet sont augmentées par les nouvelles commandes reçues et diminuées par la valeur des livraisons. Cette vérification s'assure que la valeur des commandes en carnet de la période précédente est inférieure à la somme des livraisons et des commandes en carnet pour le mois courant, assurant ainsi que les nouvelles commandes reçues sont positives.

v) Rapports cumulatifs des livraisons

Certaines entreprises rapportent les livraisons de façon cumulative au cours d'une année. Pour s'assurer que de tels cas sont corrigés, on vérifie le nombre successif d'augmentations dans les livraisons.

D. Rapports pour la gestion.

De façon à assurer une bonne gestion et un bon contrôle de l'enquête, les huit rapports suivants sont produits:

- une **liste maîtresse de l'échantillon**, contenant la liste de tous les établissements qui doivent être contactés par un bureau régional,

- une **liste de mise à la poste**, contenant la liste des établissements devant rapporter par la poste. Cette liste est aussi utilisée pour contrôler la mise à la poste et la réception des questionnaires,

- une **liste des naissances**, identifiant les nouveaux établissements qui doivent être contactés pour un mois donné,

- un **rapport sur la planification des entrevues**, permettant au gestionnaire de connaitre la liste des établissements qui doivent être contactés par téléphone, pour chaque jour de la période de collecte . Ce rapport est produit pour chacune des périodes de collecte ,

- un **rapport statistique sur les entrevues**, permettant de connaitre le nombre d'entrevues complétées par un intervieweur, l'état de chacune des unités déclarantes, ainsi que le temps moyen pour compléter la collecte d'un questionnaire. Cette information permet aussi de mesurer le fardeau de travail associé à la tâche d'entrevue,

- un **sommaire de la période de collecte** , fournissant de l'information à date sur l'état de chacune des unités déclarantes. Ce rapport peut être produit à n'importe quel moment de la période de collecte,

- un **rapport détaillé de la période de collecte** , identifiant chacun des répondants, en fonction de son état de réponse (donnée valide, en erreur ou manquante). L'objectif est de permettre au gestionnaire d'identifier des patrons de réponse d'une période de référence à une autre, et

- un **rapport mensuel des activités**, fournissant l'information nécessaire au gestionnaire d'un bureau régional pour comparer les résultats de la collecte des données de son bureau régional avec les résultats des autres bureaux régionaux.

E. Transmission électronique des données.

Tout au long de la période de collecte , les données sont transmises de façon régulière au siège social. Le logiciel déjà existant dans les bureaux régionaux est utilisé à cette fin.

3. Implantation

Le système a été développé avec les mini ordinateurs Honeywell, en utilisant les logiciels: *«Data Entry Facility (DEF)»*, *«COBOL avancé»*, *«SORT/MERGE»* et le *langage de contrôle Honeywell*. Le système a été conçu pour être facile d'opération, par l'utilisation de menus.

Une version initiale, mais complète du système IAO, a été implanté au bureau régional de Halifax en mars 1987, de façon à effectuer un essai complet du système avant son implantation dans les autres bureaux régionaux. Les principaux objectifs de cet essai étaient:

i) de vérifier le fonctionnement du système avec des données réelles,

ii) de vérifier la qualité des données résultant de l'utilisation du nouveau système,

iii) de vérifier les outils de gestion et de contrôle du système,

iv) d'obtenir la réaction des intervieweurs concernant le sytème IAO, et

v) d'identifier les points qui pourraient être améliorés avant l'implantation dans les autres bureaux régionaux.

Le test a duré 5 mois et a permis d'identifier certains problèmes et certaines lacunes du système. L'application du test a permis de modifier le système, avant son implantation dans les autres bureaux régionaux. Les principales améliorations apportées au système ont été:

i) d'améliorer le système de re-planification des entrevues, lors d'un «non-contact»,

ii) de rendre plus facile, aux intervieweurs, l'utilisation du système, le rendre plus humain,

iii) d'améliorer le traitement des non-réponses et des rapports saisonniers,

iv) de rendre plus clair et plus fonctionnel l'identification des variables qui doivent être rapportées par un répondant,

v) de permettre l'impression, dans les bureaux régionaux, des étiquettes pour le guide des répondants,

vi) de faciliter la correction des erreurs faites sur des rapports précédents, et

vii) après une certaine date, de permettre la planification automatique d'entrevues téléphoniques pour les unités déclarant par la poste.

4. Qualité des données

Afin de vérifier la qualité des données produites par le système IAO, une collecte parallèle des données a été faite avec l'ancien système. La comparaison des données de l' ancien système avec celles du nouveau a montré que ce dernier ne cause aucune détérioration des données. Les observateurs du bureau régional de Halifax croient que le nouveau système a permis d'améliorer la qualité des données, en assurant une vérification plus structurée et systématique des données.

Une étude supplémentaire, comparant le taux et la consistance des vérifications du nouveau et de l'ancien système pour les bureaux régionaux de Halifax et de Winnipeg a été réalisée.

i) Pour l'ancien système

Les intervieweurs devaient, de façon manuelle, identifier et vérifier toutes les transactions montrant un changement mensuel supérieur à 25 %. Toutes les transactions qui ont été vérifiées par un intervieweur , entre les mois d'octobre 1985 et d'avril 1986 ont été envoyées au siège social.

ii) Pour le nouveau système

Le système de détection des données aberrantes qui devait être implanté dans les bureaux régionaux a été appliqué aux données initiales pour la même période de référence.

Le taux de rejet, ainsi que la variation de ce taux pour différents mois ont été calculés pour les deux systèmes. De plus, pour l'ancien système, le taux espéré de rejet pour l'ensemble de la période a aussi été calculé. Des résultats partiels pour la caractéristique Livraisons sont présentés dans la table suivante:

TAUX DE REJET OBSERVÉ

MOIS DE RÉFÉRENCE	HALIFAX		WINNIPEG	
	ANCIEN	NOUVEAU	ANCIEN	NOUVEAU
Oct. 85	26.4	16.2	31.8	11.2
Nov. 85	32.7	12.7	37.4	12.7
Dec. 85	33.7	18.3	42.4	13.4
Jan. 86	41.0	16.1	47.9	11.0
Fev. 86	--	--	45.0	16.6
Mar. 86	--	--	42.8	19.1
Avr. 86	--	--	46.2	14.2
TAUX MOYEN	33.2	15.8	41.9	14.1
TAUX ESPÉRÉ	52.0	--	52.0	--

NOTE: Voir C.S.I.O. REGIONALIZATION - PRELIMINARY REPORT, pour les résultats détaillés de cette étude.

En analysant ces résultats, on remarque que le taux de rejet pour l'ancien système varie beaucoup plus d'un mois à l'autre que celui du nouveau système. Le nouveau système semble s'adapter beaucoup mieux à la saisonnalité des données que la règle arbitraire qui consistait à rejeter les unités ayant un changement supérieur a 25 % en valeur absolue. On note aussi que le taux espéré de rejet pour l'ancien système est supérieur au taux de rejet observé par les intervieweurs. Cette situation peut avoir un effet significatif sur la qualité des données produites par l'ancien système. En effet, certaines unités qui n'ont pas été vérifiées par les intervieweurs, mais qui auraient du l'être, peuvent être en erreur.

On note aussi que le nouveau système de détection des données aberrantes identifie beaucoup moins d'unités. De façon à s'assurer que le nouveau système identifie bien les erreurs, une comparaison entre les unités rejetées par ce système et les unités qui ont été modifiées suite à une vérification des intervieweurs sous l'ancien système a été effectuée pour la variable Livraisons. Sur les 131 corrections apportées par les intervieweurs, 129 ont été identifiées par le nouveau système. On remarque donc que le nouveau système, tout en rejetant moins d'unités pour la vérification, identifie pour ainsi dire, toutes les transactions corrigées par les intervieweurs. De plus, le nouveau système étant appliqué de façon automatique et uniforme, assure un meilleur contrôle sur la qualité globale des données.

5. Conclusion

Le système d'Interview Assisté par Ordinateur pour l'Enquête sur les Stocks, Livraisons et Commandes des industries manufacturières représente une amélioration certaine sur l'ancien système de collecte et de saisie des données. Le nouveau système utilise avantageusement l'ordinateur pour améliorer la qualité des données, réduire le fardeau de réponse et diminuer les coûts associés à la collecte de l'information.

Le module de planification automatique des entrevues permet de s'adapter à la disponibilité des répondants. Le système d'interview interactif permet de réduire le nombre de rappels. Le système de détection des données aberrantes utilisé assure une meilleure qualité globale de l'information, effectuant un contrôle structuré, automatique et uniforme. Les rapports pour la gestion permettent un meilleur contrôle des opérations.

6. Bibliographie

CSIO - Regionalization working group, *CSIO regionalization preliminary report*, rapport interne, Statistique Canada, décembre 1986.

Tambay J.-L., *Study of outliers in the CSIO regional offices*, rapport interne, Statistique Canada, mars 1986.

Tambay J.-L., *Application of tolerance edits to regional office corrections*, rapport interne, Statistique Canada, juin 1986.

ENQUÊTE SANTÉ-QUÉBEC

Plan de sondage

Présentation faite au
Colloque sur les applicatons de la statistique
dans le cadre du
cinquante-cinquième congrès pour
l'avancement des sciences

le 19 mai 1987
à l'université d'Ottawa

Robert Courtemanche
Françoise Tarte
Direction de la méthodologie et de l'information
Bureau de la statistique du Québec

1. PRÉSENTATION DE L'ENQUÊTE SANTÉ-QUÉBEC

Le ministère de la Santé et des Services sociaux (MSSS) conjointement avec les trente-deux départements de santé communautaire (DSC), suite à la réalisation de l'enquête Santé-Canada et à cause de la faible représentativité de l'échantillon de cette enquête au niveau des grandes régions du Québec, ont décidé de mettre sur pied une enquête sur la santé des québécois, l'enquête Santé-Québec.

Le projet d'enquête est en cours depuis 1980. Au printemps 1985, les responsables du dossier au MSSS ont fait appel aux méthodologistes du BSQ pour qu'ils développent et mettent sur pied le plan de sondage et le système d'estimation de cette enquête.

Un des principaux objectifs de l'enquête Santé-Québec est de satisfaire les besoins en données sur la santé physique et mentale des Québécois à partir de la population du Québec et non pas de ses institutions. Ces renseignements permettront aux planificateurs et aux décideurs d'avoir une vision plus globale de la santé et plus éclairée. De plus, l'enquête Santé-Québec vise à avoir une représentation régionale satisfaisante.

2. CONCEPTION DES QUESTIONNAIRES

Les questionnaires de l'enquête Santé-Québec, le QRI (questionnaire rempli par l'interviewer) et le QAA (questionnaire auto-administré) ont été conçus à partir des questionnaires de l'ESC et de d'autres enquêtes de santé réalisées dans des contextes comparables à ceux de l'ESQ et dont l'enquête pilote tirait son contenu. Ceci permettait d'assurer une certaine comparabilité des questions.

Le QRI est un questionnaire qui s'adresse aux personnes du ménage. Le QAA est un questionnaire laissé au ménage que chaque personne de 15 ans et plus doit remplir et retourner à la firme privée.

Les principaux thèmes abordés sont les habitudes de vie (consommation d'alcool, usage du tabac et de drogues, activités physiques, usage de véhicules-moteurs, prévention féminine), les problèmes de santé physique (limitations d'activité, problèmes de courtes durées, problèmes chroniques, auditions, vision, accidents, blessures), la consommation des services et des soins (les professionnels rencontrés, le lieu de traitement, l'usage de médicaments), les antécédents (personnels et familiaux), l'environnement (travail, revenu, âge, sexe, scolarité), les problèmes de santé psychologique (suicide, échelle IlFeld, échelle Dupuy, événements stressants) et de santé sociale (soutien social, problèmes sociaux) et les conséquences sur la fonctionnalité (journées d'incapacité, mobilité, besoin d'aide).

3. POPULATIONS ET UNITÉS D'ENQUÊTE

3.1 La population visée

La population visée est l'ensemble des ménages privés de toutes les régions socio-sanitaires du Québec à l'exclusion de la région socio- sanitaire 10 (Nouveau-Québec) et des réserves indiennes. Une région socio-sanitaire (RSS) est composée de territoires de DSC. Il y a entre 1 et 8 territoires de DSC par RSS.

Un ménage privé est, d'après la définition du recensement canadien "la ou les personnes (autres que des résidents étrangers) qui occupent un logement privé et qui n'ont pas de lieu de résidence habituelle ailleurs au Canada [1]. "Un logement privé est" un ensemble structuralement distinct de pièces d'habitation de d'autres logements ayant une entrée privée donnant sur l'extérieur ou sur un corridor ou un escalier commun à l'intérieur, c'est-à-dire que l'entrée doit donner accès au logement sans qu'on ait à passer par les pièces d'habitation de quelqu'un d'autre" [4].

La région du Nouveau-Québec a été éliminée à cause des problèmes de langue (on n'y parle ni français, ni anglais) et des coûts de déplacements élevés. La région 10A et le DSC 102 représentent le territoire Inouit, la région 10B et le DSC 101 représentent le territoire CRI. Cette population exclue représente environ 0,09% de la population totale des ménages privés du Québec. Les réserves indiennes sont éliminées à cause des difficultés potentielles d'accès sur ces territoires: en effet, chaque chef de bande de chaque réserve a le droit d'autoriser ou non la présence d'interviewers sur son territoire. Nous ne pouvions pas nous permettre d'avoir seulement une partie de ces réserves pour établir un plan de sondage. De plus, ces groupes sont déjà des sujets d'études dans certaines universités et ils représentent des caractéristiques propres relativement à leur mode de nutrition et à leur état de santé. De plus, certaines parties des soins de première ligne sont financées par le gouvernement fédéral. Finalement, l'enquête Santé-Canada les avait exclues.

C'est ainsi que 5 réserves indiennes, représentant 0,28% de la population des ménages privés du Québec ont été éliminées. Elles sont Whitworth, Cacouna, Coucouache, St-Augustin, Kawawachikamach.

Nous laissons de côté en plus de la région 10 et des réserves indiennes, les ménages collectifs qui regroupent une population généralement plus âgée que celle des ménages privés. Les ménages collectifs représentent 0,28% de l'ensemble des ménages du Québec.

3.2 La population échantillonnée

Il s'agit de la population visée moins les territoires non organisés, moins les secteurs de dénombrement de cinq ménages et moins, moins les villes mortes.

Les territoires non organisés sont des régions sans organisation municipale administrées directement par la province ou par le comté. Elles sont caractérisées par la faible densité de leur population et l'étendue de leur territoire. Ces régions représentent environ 0,11% de la population totale des ménages privés du Québec.

Les secteurs de dénombrement sont les unités de base de toutes les unités géostatistiques du recensement. Ceux qui comprennent moins de 5 ménages selon les données du recensement de 1981 constituent environ 0,08% de la population totale des ménages privés du Québec. Ce sont également des secteurs couvrant de grands territoires.

Les villes mortes depuis 1981 sont Gagnon et Schefferville; ceci représente environ 0,08% de la population totale des ménages privés. Cette étape constitue une mise à jour des données de 1981.

La population échantillonnée représente environ 99,4% de la population visée, et 98,5% de l'ensemble des ménages du Québec.

3.3 L'unité primaire d'échantillonnage

L'unité primaire d'échantillonnage est formée d'aires géographiques. La section **base de sondage** décrit clairement la composition des UPE. Il y a entre 77 et 1671 UPE par DSC. Le plan de sondage de l'enquête Santé-Québec est un plan de sondage à deux degrés. Au premier degré, on tire des unités primaires d'échantillonnage dans chacun des DSC. Au deuxième degré, on tire des unités secondaires d'échantillonnage.

3.4 L'unité secondaire d'échantillonage

L'unité secondaire d'échantillonnage de l'enquête Santé-Québec est le logement privé tel que décrit précédemment. Pour chaque aire géographique échantillonnée, des énumérateurs ont été formés en vue de recenser tous

les logements privés de ce secteur c'est-à-dire les maisons privées, double ou en rangée, les duplex, triplex ou autre types d'appartement, les roulottes, etc. Les logements exclus sont les hôtels, motels ou maisons de chambres, les résidences scolaires, les foyers et centres d'accueil, les terrains de camping, les établissements pour personnes âgées, handicapés et malades chroniques, les pensions, etc.

3.5 L'unité enquêtée

L'unité enquêtée de l'enquête Santé-Québec change selon le type de questionnaire à remplir. Pour la fiche d'identification du ménage et pour le QRI tous les membres du ménage choisi constituent les unités à enquêter. Un membre du ménage est une personne qui considère le logement qu'elle occupe comme son lieu de résidence habituel ou qui habite le logement et qui n'en a pas d'autres ailleurs [4]. Pour le QAA, tout membre du ménage de 15 ans et plus admissible au QRI peut remplir le questionnaire à l'exception de ceux dont l'incapacité est reliée à leur difficulté à comprendre et à suivre un questionnaire écrit ou un questionnaire lu par une tierce personne.

3.6 L'unité répondante

Pour le QRI, l'unité répondante du ménage est toute personne de dix-huit ans et plus du ménage apte à répondre pour tous les autres membres du ménage quel que soit leur âge. Elle répondra donc pour tous les membres du ménage. Pour le QAA, l'unité répondante est l'unité enquêtée.

4. OPÉRATIONS DE TERRAIN

Les opérations de terrain comprennent deux phases distinctes: l'énumération des logements des UPE et la collecte des données. Les opérations de terrain se font en huit périodes étalées sur une année complète.

4.1 L'énumération

Chaque période d'énumération se fait sur un maximum de 4 jours et précède la période de collecte correspondante d'au moins 1 mois et demi. L'énumération consiste à dresser la liste de tous les logements privés se situant à l'intérieur des limites d'une UPE afin de recencer et de localiser tous les logements privés d'une UPE.

- pour tirer l'échantillon de logements;
- pour favoriser les entrevues dans les logements difficiles à repérer;
- pour faire de meilleures estimations.

Les logements qui ne doivent pas être inscrits sur ces listes sont les logements saisonniers, les logements collectifs, les institutions, les commerces, etc. Les logements vacants doivent être énumérés car ils pourront être occupés lors des interviews.

Les énumérateurs qui travaillent dans les UPE formées à partir de côtés d'îlots ont une nouvelle UPE à chaque période; les autres ont 4 UPE à énumérer, l'UPE formée à partir de SD servant pour deux périodes d'entrevues (la première et la cinquième, la deuxième et la sixième et ainsi de suite).

4.2 La collecte des données

Chaque période de collecte de données s'étale sur trois semaines; elles sont séparées les unes des autres par environ 1 mois et demi. Les huit périodes assurent une charge annuelle aux enquêteurs, ce qui aide à garder les mêmes et à augmenter la qualité des entrevues d'où une meilleure qualité des données. Pour un même nombre total d'entrevues à faire, les coûts de recrutement et d'entraînement sont réduits si on étale l'enquête sur une année plutôt qu'en un ou deux terrains car le nombre d'interviewers à former est moins grand. D'un point de vue méthodologique, le choix de la période de collecte est moins critique lorsqu'il y en a plu-

sieurs; les estimations annuelles pour les variables ayant des variations saisonnières (consommation d'alcool, activités physiques extérieures, etc) seraient de meilleures qualités.

Huit périodes ont été fixées plutôt que dix ou douze pour empêcher les différentes opérations reliées aux périodes de collecte de se chevaucher et pour éliminer les périodes difficiles à enquêter telles le temps des Fêtes, Pâques et les périodes de vacances.

L'échantillon est réparti de façon égale à chaque période de collecte. Il y a 1720 logements échantillonnés à chaque collecte. Chaque enquêteur a une liste de 10 logements pointés sur la liste des logements énumérés. La majorité des entrevues seront faites au cours des deux premières semaines, la troisième semaine étant plutôt réservée aux rappels pour les non-répondants. La tâche de 10 logements à enquêter par interviewer par période de collecte a été fixée d'après la durée des entrevues de l'enquête pilote les distances à parcourir durant les semaines d'enquête et les possibilités de contacts des ménages.

5. BASE DE SONDAGE

L'enquête Santé-Québec repose sur une base de sondage aréolaire. Les aires sont appelées unités primaires d'échantillonnage (UPE). La liste complète des UPE avec ses caractéristiques constitue la base de sondage soit la population échantillonnée. Les principales caractéristiques sont la localisation géographique et le nombre de ménages privés résidant dans l'UPE selon le recensement de 1981, (le nombre de ménages étant utilisé pour le tirage des UPE, il a été mis à 0 pour les villes mortes décelées depuis 1981 comme Gagnon et Shefferville).

Deux types d'UPE ont été formées selon la nature des données que nous avions en main. Dans les départements de santé communautaire (DSC) qui étaient inclus intégralement dans des régions métropolitaines de recensement soient les DSC de l'Ile de Montréal et le DSC St-Sacrement, les données sur la population de 1981 sont connues jusqu'au niveau du côté d'îlot. Dans les autres DSC, la plus petite unité géostatistique correspond au secteur de dénombrement (SD). Ainsi les UPE sont formées à partir des plus petites unités géostatistiques disponibles soient les côtés d'îlot ou les SD.

5.1 UPE formées à partir de SD

Le nombre de ménages privés résidant dans un secteur de dénombrement varie entre 0 et 600 ménages selon le recensement canadien de 1981. Pour homogénéiser la taille des UPE les SD ont été regroupés de façon à obtenir un minimum de 80 ménages par UPE. Ce nombre est un compromis entre l'importance de la tâche à faire lors de l'énumération et l'effet sur la qualité du plan pour assurer une certaine dispersion (géographique) de l'échantillon.

Le regroupement de SD (généralement deux SD servaient à former une UPE lorsqu'un regroupement était nécessaire) s'est fait selon certains critères prioritaires et certains critères secondaires.

Le premier critère prioritaire - le nombre total de ménages des SD regroupés doit être plus grand ou égal à 80 - assurait une taille minimale aux UPE formées pour permettre une certaine flexibilité dans le tirage de l'échantillon et une certaine dispersion géographique de l'échantillon.

Le second critère prioritaire - l'existence d'un chemin réunissant les SD regroupés - réduisait les déplacements des énumérateurs - interviewers et les coûts reliés à ces derniers.

Le troisième critère prioritaire - les SD regroupés doivent appartenir à une même strate - facilitait le traitement opérationnel (statistique) de ces regroupements.

Les critères secondaires avaient pour buts principaux d'uniformiser la tâche des énumérateurs-interviewers en fonction du nombre de ménages et des déplacements et de diversifier la composition des ménages de l'UPE formée:

- choisir, pour un secteur donné, parmi les regroupements possibles celui qui minimise (uniformise) le nombre de ménages du regroupement
- choisir, pour un secteur donné, parmi les regroupements possibles celui qui minimise (uniformise) la grandeur du territoire du regroupement
- choisir, pour un secteur donné, parmi les regroupements possibles celui qui rend l'UPE formée la plus hétérogène selon certaines variables socio-économiques.

Les UPE formées à partir de SD correspondent dans 95% des cas à des SD uniques et les autres UPE sont des regroupements de SD. Leur taille varie entre 80 et 575 ménages privés avec une moyenne de 242. Même si nous avons tenté d'uniformiser le nombre de ménages, comme nous le propose Kish en [3], il reste cependant que les UPE ainsi formées ne le sont pas parfaitement et que, pour obtenir plus d'homogénéité, il aurait fallu scinder des SD après l'énumération ce qui s'avérait trop coûteux et impliquait des délais opérationnels que nous ne pouvions pas nous permettre.

5.2 UPE formées à partir de côtés d'îlots

Les DSC de l'Île de Montréal ainsi que le DSC St-Sacrement de Québec sont découpés en UPE qui sont formées à partir de côtés d'îlots. Un côté d'îlot se définit comme une arête d'un pâté de maisons. Le nombre de ménages privés résidant dans un côté d'îlot varie entre 0 et près de 400 ménages

selon le recensement canadien de 1981. Dans le but d'uniformiser la taille des UPE les territoires ont été construits selon cinq types de composition d'unités géographiques:

1. UPE formée d'un côté d'îlot
2. UPE formée d'un îlot
3. UPE formée d'un regroupement de côtés d'îlots et d'îlots inclus dans un secteur de dénombrement
4. UPE formée d'un secteur de dénombrement
5. UPE formée d'un regroupement de secteurs de dénombrement

Les mêmes critères que ceux utilisés pour la formation des UPE à partir des SD ont été repris pour la formation des UPE à partir des côtés d'îlots à l'exception que le nombre minimal de ménages privés a été fixé à 40. Les principales étapes pour la formation des UPE formées à partir des côtés d'îlots ont été les suivantes:

- Les SD ayant moins de 40 ménages privés ont été regroupés avec un autre secteur de dénombrement contigüe selon les critères déjà mentionnés pour former le cinquième type de composition d'UPE
- Si un îlot est formé de côtés d'îlot ayant tous plus de 40 ménages privés, chaque îlot forme une UPE de type 1
- Si chaque îlot à l'intérieur d'un secteur de dénombrement comprend au moins 40 ménages privés, chaque îlot forme une UPE de type 2
- Les SD ayant moins de 200 ménages privés et n'étant pas formés d'UPE sont automatiquement classés comme UPE de type 4.
- Pour les autres SD , divers algorithmes ont été développés pour former des UPE de type 3.
- Finalement les SD où aucun algorithme n'a fonctionné ont été classés de type 4.

La distribution des UPE formées à partir de côtés d'îlots, selon leur type, est détaillée au tableau 1 (voir annexe). Comme nous pouvons le constater les divers algorithmes nous ont permis de subdiviser la majorité des secteurs de dénombrement et ainsi de diminuer la tâche de l'énumération et la variation du nombre de ménages entre les UPE. Il a été cependant impossible de tenir compte de diverses variables socio-économiques pour s'assurer de l'hétérogénéité à l'intérieur des UPE. Cette lacune a été contrebalancée par le grand nombre d'UPE échantillonnées (voir tableau 2 en annexe).

6. STRATIFICATION

La stratification de l'enquête Santé-Québec devait tenir compte des territoires des 32 départements de santé communautaire. La demande du MSSS était formelle, il fallait s'assurer d'une taille d'échantillon fixe par DSC. Il a alors été envisagé de partitionner les DSC en 2 strates: une urbaine et l'autre rurale.

Plusieurs possibilités ont alors été étudiées pour mettre en place cette stratification rurale-urbaine:

1. Utiliser un plan de sondage à 3 degrés dans les DSC mixtes:
   - 1er degré: regroupement de plusieurs SD de même type (rural ou urbain) correspondant à une tâche annuelle d'interviewer
   - 2e degré: les UPE conformes à ceux déjà décrits
   - 3e degré: les ménages
2. Utiliser un plan de sondage à deux degrés où au
   - 1er degré: les UPE conformes à ceux déjà décrits. Les UPE sont ensuite regroupées ensemble selon leurs caractéristiques rurales ou urbaines pour former des tâches d'interviewer.
   - 2e degré: les ménages

L'hypothèse (1) a été rejetée due aux contraintes de temps que nous avions et d'un manque de ressources méthodologiques et opérationnelles.

L'hypothèse (2) a été rejetée puisque les regroupements d'UPE de même type ainsi formés occasionnaient des déplacements supplémentaires considérables des interviewers par rapport aux regroupements d'UPE sans la composante rurale urbaine. Aucune partition n'a donc été faite dans les DSC. Les strates de l'enquête correspondent donc aux 32 territoires de DSC. Les limites des territoires des DSC utilisés proviennent du système J36 du ministère de la Santé et des services sociaux d'août 1985. Ce découpage est caractérisé entre autres, par la réallocation du district Mont-Laurier du CLSC des Hautes Laurentides de la région de l'Outaouais à la région Laurentides Lanaudière.

7. TAILLE ET RÉPARTITION DE L'ÉCHANTILLON

En mai 1985, quand le ministère de la Santé et des services sociaux a demandé au BSQ de collaborer dans le projet, la taille totale de l'échantillon et sa répartition étaient à peu près fixées. Les règles étaient les suivantes:

- 300 ménages minimum échantillonnés par DSC
- 900 ménages minimum échantillonnés par RSS
- 12900 ménages échantillonnés au total

Les buts à atteindre visés à travers les règles, étaient de pouvoir établir un certain portrait de la population au niveau des DSC, de comparer les données au niveau des RSS et de pouvoir avoir une très bonne précision sur des estimations de cas assez rares au niveau du Québec.

L'échantillon au niveau des DSC n'était pas proportionnel à la taille du DSC mais fixe étant donné que chaque DSC participait de façon égale aux composantes de l'enquête. Compte tenu de la grande taille de population de chaque DSC, les tailles de leur échantillon pour obtenir un niveau de qualité fixé ne sont pas influencées par elles.

Divers remaniements mineurs ont été apportés à ces règles de départ sans toutefois changer les buts à atteindre:

- Une augmentation des tailles minima d'échantillon de l'ordre de 10% a été accordée à chaque DSC pour tenir compte des hors-échantillons non décelés à l'énumération tels les logements vacants, les logements collectifs (composés de plus de 10 personnes non apparentées à la personne de référence etc.). Ce taux de 10% correspond au taux de hors-échantillon de l'enquête sur la population active du Canada.

- Une réallocation du nombre de ménages dans les régions du Nord-Ouest et de la Côte-Nord a été redistribué à trois DSC de la région de Montréal (Ste-Justine, Montréal Général, St-Luc) pour compenser les pertes du nombre de personnes dues au faible ratio du nombre moyen de personnes par ménage pour ces zones très fortement urbanisées où en plus il est prévu que les taux de réponses sont plus faibles dans ces régions.

- Une répartition de façon proportionnelle à la racine carrée du nombre de ménages a été faite dans les régions du Bas-Saint-Laurent / Gaspésie / Iles de la Madeleine et du Saguenay / Lac St-Jean où il n'y avait que 2 DSC. Cette répartition était un compromis d'une part entre avoir la meilleure précision au niveau de chaque DSC si l'option d'égalité de taille était choisie pour chaque DSC et d'autre part, avoir la meilleure précision au niveau de la région socio-sanitaire si l'option de proportionnalité à la taille de chaque DSC avait été choisie.

- Une répartition telle que le nombre de ménages par DSC soit divisible par 80 pour établir les tâches d'interviewer c'est-à-dire 10 entrevues par période par DSC. Cette répartition nous permet d'obtenir un échantillon balancé chronologiquement.

La répartition de l'échantillon est donc la suivante:

- 320 logements échantillonnés au minimum par DSC
- 960 logements échantillonnés au minimum par RSS sauf pour les régions 8 et 9
- 13760 logements échantillonnés au total

Il fallait par la suite choisir la répartition de ces logements à l'intérieur des strates tout en respectant la tâche d'un interviewer. Dans les cas où les UPE étaient formées de côtés d'îlots, on a choisi de prendre 10 ménages par UPE, ces UPE étant généralement plus petites. Dans les autres cas, on a choisi d'échantillonner 20 ménages par UPE pour diminuer le travail d'énumération.

8. TIRAGE DE L'ÉCHANTILLON

Le tirage de l'échantillon se fait en plusieurs étapes: tirage des UPE, répartition dans le temps des UPE, énumération des UPE, tirage des logements dans les UPE.

Le tirage des UPE a débuté par un triage des UPE selon leur caractéristique rurale et urbaine. Les UPE ont été tirées systématiquement avec probabilité proportionnelle à leur nombre de ménages en 1981 pour que les UPE plus populeuses aient plus de chances d'être choisies.

Les UPE échantillonnées ont été réunies en tâche d'interviewer: soit 8 UPE (1 UPE par période) pour les UPE formées à partir de côtés d'îlots ou soit 4 UPE (1 UPE pour 2 périodes) pour les UPE formées à partir de SD. Le regroupement manuel s'est fait de manière à minimiser les distances entre les UPE réunies et ainsi minimiser les déplacements des interviewers. Ces regroupements, appelés unités de calcul, servent pour la mesure de variance globale incluant la variance due à la méthode d'échantillonnage et la variance des interviewers.

À l'intérieur de chaque groupe d'UPE réunies en tâches, la période de collecte a été assignée de façon aléatoire.

Puis l'énumération a permis de dresser la liste des logements privés des UPE échantillonnées. Le tirage des logements se fait suite à l'énumération. Les logements sont choisis systématiquement avec un pas fractionnaire à partir de ces listes. Contrairement à d'autres enquêtes, le pas d'échantillonnage dépend du nombre de ménages énumérés. Ceci était dû au fait qu'il fallait avoir une taille d'échantillon fixe pour que la firme puisse mieux évaluer les coûts de terrain. Le pas consiste à diviser le nombre de logements énumérées par la taille d'échantillon.

CONCLUSION

Cette expérience fort complexe qu'est l'enquête Santé-Québec est l'aboutissement d'un effort conjoint de collaboration du secteur privé et gouvernemental.

C'est la première fois qu'au Québec et probablement même au Canada des structures fonctionnelles aussi différentes se réunissaient pour une enquête d'une telle envergure.

Pour la firme, les contraintes méthodologiques du BSQ se traduisaient souvent par de nouvelles méthodes et approches. Cela impliquait également la gestion et l'encadrement d'une équipe de plus de 200 personnes.

Pour le BSQ, cela impliquait de développper pour la première fois toute l'expertise d'une enquête par interviewer auprès des ménages: établir un plan d'échantillonnage à plusieurs degrés sans base préalable, former la base de sondage et toute la programmation sous-jacente, déterminer les procédures de tirage, le tirage et la programmation, établir les populations cibles et échantillonnées, élaborer une méthode d'énumération et l'enseigner, apprendre des concepts divers concernant la santé, les individus, les ménages, les logements, les unités géostatistiques. Finalement, agir comme consultant à plusieurs niveaux.

ANNEXE

Tableau 1

Distribution des UPE formées à partir de côtés d'îlots

Type	Nombre d'UPE	Proportion
1	700	4,4
2	5 196	54,9
3	3 498	36,9
4	54	0,6
5	19	0,2

Tableau 2

Diverses caractéristiques des DSC

Nom du DSC	Numéro de DSC	Nombre de ménages privés (population échantillonnée)	Nombre de logements à échantillonner	Nombre d'UPE	Nombre d'UPE échantillonnées	Nombre moyen de personnes par ménage
Rimouski	11	44 860	560	219	28	3,24
Gaspé	12	22 360	400	126	20	3,68
Chicoutimi	21	49 805	480	155	24	3,37
Roberval	22	35 630	480	164	24	3,50
Beauceville	31	45 345	320	217	16	3,26
Rivière du loup	32	27 670	320	145	16	3,27
Québec	33	62 000	320	225	16	3,21
Lévis	34	38 550	320	147	16	3,15
Chul (Québec)	35	76 715	320	300	16	3,08
St-Sacrement (Québec)	36	62 404	320	860	32	2,36
Montmagny	37	14 120	320	77	16	3,31
Shawinigan	41	32 640	320	122	16	2,94
Drummondville	42	49 340	320	207	16	3,07
Trois-Rivières	43	64 210	320	270	16	2,94
Sherbrooke	51	85 485	960	363	48	2,91
Lakeshore	61	73 520	320	1 016	32	2,91
Verdun	62	71 675	320	723	32	2,55
Maisonneuve-Rosemont	63	149 272	320	1 671	32	2,64
Ste-Justine	64	70 134	400	689	40	2,30
Montréal Général	65	87 566	400	876	40	2,16
Sacré-Coeur (Cartierville)	66	126 353	320	1 401	32	2,63
Saint-Luc	67	110 324	400	994	40	2,25
Laval	68	85 132	320	1 237	32	3,10
St-Jérome	161	99 650	480	427	24	3,01
Joliette	162	78 995	480	340	24	3,16
Valleyfield	261	62 165	320	256	16	3,02
St-Hyacinthe	262	94 160	320	369	16	3,04
Greenfield Park	263	129 150	320	415	16	3,02
St-Jean	264	55 820	320	250	16	2,99
Hull	071	78 505	960	335	48	3,03
Rouyn	081	45 455	880	195	44	3,18
Hauterive	091	31 365	800	147	40	3,43

BIBLIOGRAPHIE

1 Cox Brenda G, Cohen Steven B., **Methodological issues for health care surveys**, D.B., Owen, University of Michigan, 446p.

2 Hansen, Huwitz, Madow, **Sample Survey Methods and Theory**, Volume 1, Wily Publications in Statistis, 638p.

3 Kish, Leslie, Survey Sampling, John Wiley and Sons, Inc. 643p.

4 Statistique Canada, **Dictionnaire du recensement de 1981**, catalogue 99-901.

FORMALISATION DU CALCUL DES TAUX DE RÉPONSE EN PRÉSENCE D'UNITÉS HORS POPULATION

MICHEL FLUET[1]

Le présent article se veut un document de réflexion sur le calcul du taux de réponse en présence d'unités hors population. On y aborde successivement la perception du taux de réponse comme mesure de qualité. Différentes méthodes pour mesurer ce taux ainsi que différents cas d'inéligibilité et la façon de les intégrer au calcul du taux de réponse.

1. LA PERCEPTION DU TAUX DE RÉPONSE EN TANT QUE MESURE DE LA QUALITÉ

Dans le milieu des sondages ou des enquêtes, que ce soit au niveau gouvernemental ou dans l'entreprise privée, le taux de réponse est perçu, à tort ou à raison, comme une des mesures les plus importantes pour juger de la qualité d'un sondage.

Cette mesure a pris tellement d'importance aux yeux des entreprises privées commandant des enquêtes que le taux de réponse est devenu un élément de décision de plus en plus présent lorsqu'une entreprise choisit de retenir les services de firmes spécialisées pour réaliser un sondage, une étude de marché ou encore une cueillette de données. En outre, certains médias hésitent à publier les résultats de sondages en invoquant tout simplement que les taux de réponse sont trop faibles!!!

Devant l'importance accrue imputée aux taux de réponse et en l'absence de normes établies et reconnues, les firmes spécialisées en recherche, sondage ou marketing ont tendance à multiplier les méthodes pour calculer les taux de réponse associés à leurs résultats. Cette absence de normes rend la comparaison des taux de réponse très difficile, ce qui nuit considérablement à la crédibilité des firmes spécialisées dont les sondages font souvent l'objet de comparaisons. Il incombe alors à tous ceux et celles oeuvrant dans ce milieu de se doter de normes qui seront reconnues tant par les clients que par les spécialistes du domaine des sondages et des enquêtes.

De plus, il est important de faire saisir à tous les clients que le taux de réponse n'est

1. Michel Fluet est directeur de la maison de recherche SOM à Montréal.

pas la seule mesure de la qualité d'un sondage. En fait, le taux de réponse est l'une des mesures utilisées pour évaluer la qualité d'une cueillette des données et à ce titre, il doit être perçu comme un indicateur de la qualité associée à un sondage. De fait, il n'est que l'une des mesures associées à la qualité d'un sondage.

2. LE CALCUL DU TAUX DE RÉPONSE

Une des premières étapes à franchir pour se diriger vers une normalisation du calcul d'un taux de réponse serait d'établir une liste exhaustive de toutes les méthodes actuellement utilisées pour le calculer. Il serait toutefois difficile, voire superflu, de vouloir dresser cette liste afin d'analyser et de comprendre la logique qui est à la base de chaque méthode pour ainsi dégager toute l'ampleur des différences que l'on pourrait rencontrer. Les deux formules ci-dessous illustrent deux méthodes diamétralement opposées qui fournissent un très bon exemple des différences qu'il peut exister.

Formule 1: Le nombre d'entrevues complétées éligibles en pourcentage du nombre total d'unités.

$$\frac{\text{Nombre d'entrevues complétées éligibles}}{\text{Échantillon total}} \quad (1)$$

Formule 2: Le nombre d'entrevues complétées éligibles en pourcentage du nombre de "contacts" effectués.

$$\frac{\text{Nombre d'entrevues complétées éligibles}}{\text{Nombre de "contacts"}} \quad (2)$$

La formule 1 fournit le taux de réponse le plus bas, l'échantillon total étant le dénominateur le plus élevé. Ce dernier inclut les unités éligibles, les unités hors population (inéligibles) et les unités non rejointes. Cette méthode mesure la capacité de compléter des entrevues à l'intérieur d'un échantillon initial donné. La formule 2 fournit un taux de réponse généralement très élevé. Le nombre de "contacts" exclut généralement les unités hors population. Il inclut toutefois les entrevues complétées éligibles, les refus du répondant potentiel et les entrevues partiellement complétées et éligibles. Cette méthode mesure la capacité des interviewers à compléter des entrevues lorsqu'ils ont rejoint un répondant potentiel.

On ne peut nier que ces deux formules fournissent des indicateurs de qualité d'une cueillette de données. D'une part, à partir d'un échantillon initial, la première formule permet d'avoir une vue d'ensemble de la qualité de la base de sondage et de l'effort de cueillette. D'autre part, la deuxième formule permet d'avoir une mesure de la qualité du groupe d'interviewers concernant leur capacité à convaincre les répondants potentiels rejoints de compléter un questionnaire. Peut-on qualifier ces deux mesures de taux de réponse? On ne peut répondre oui à cette question si l'on travaille à une normalisation du calcul. Toutefois, ceci n'entraine pas le rejet automatique de ces formules avant d'engager les discussions sur la normalisation. D'ailleurs, on voit clairement que les spécialistes devront d'abord définir ce qu'ils entendent par "taux de réponse".

En théorie, le taux de réponse devrait être défini comme le nombre de personnes appartenant à la population visée ayant complété un questionnaire en pourcentage du nombre de celles qui ont été échantillonnées et qui font partie de la population visée. En pratique, ce taux est très facile à calculer si le sondage est effectué à partir d'une base où toutes les unités appartiennent à la population visée. Il en est tout autrement lorsque la base contient des éléments qui ne font pas partie de cette population.

3. L'INÉLIGIBILITÉ DANS LE CALCUL DU TAUX DE RÉPONSE

Les unités hors population sont celles qui causent le plus de problèmes dans la normalisation du calcul des taux de réponse. Une unité est déclarée hors population (inéligible) si celle-ci ne répond pas aux critères de définition de la population visée par le sondage. L'inéligibilité est la conséquence du choix d'une base de sondage qui n'est pas à jour, qui renferme des erreurs ou qui offre un contenu qui déborde le cadre de la population visée.

L'inéligibilité d'une unité peut être déclarée à plusieurs étapes du sondage:

i) <u>Contrôle de l'échantillon</u>

On identifie directement les unités hors population à partir des listes de l'échantillon. Nommons-les unités hors population inutilisables.

ii) Cueillette

À l'étape du contact, on identifie les logements démolis ou désaffectés, les logements saisonniers, les numéros de téléphone inexistants, discontinués, non-résidentiels, les faillites ou autres. Nommons-les aussi unités hors population inutilisables.

À l'étape de l'entrevue, on identifie les unités qui ne répondent pas aux critères d'appartenance à la population cible. Nommons-les unités hors population utilisables.

iii) Validation

À cette étape, on identifie les unités hors population qui ont complété un questionnaire par erreur. Nommons-les aussi unités hors population utilisables.

La méthode la plus répandue pour calculer le taux de réponse en présence d'unités hors population consiste à établir le pourcentage d'entrevues complétées éligibles en fonction du nombre total de répondants potentiels. Pratiquement, le calcul est donné par:

Formule 3:

$$\frac{\text{Nombre d'entrevues complétées éligibles}}{\text{Échantillon total - Nombre d'unités hors population}} \qquad (3)$$

Le dénominateur de cette formule contient toutefois un certain nombre d'unités hors population qui n'ont pas pu être découvertes en raison de l'incapacité de rejoindre une unité d'échantillonnage, de refus ou de l'absence prolongée du répondant potentiel.

Tenant compte de ces observations, nous proposons donc la formule suivante pour calculer un taux de réponse:

Formule 4:

$$\frac{\text{Nombre d'entrevues complétées éligibles + Nombre d'entrevues complétées inéligibles}}{\text{Échantillon total - Nombre d'unités inutilisables rejointes - Facteur de correction}} \qquad (4)$$

où le nombre d'unités inutilisables rejointes inclut le nombre d'unités hors population

inutilisables découvertes aux étapes du contrôle de l'échantillon et du contact lors de la cueillette et où le facteur de correction est une estimation du nombre d'unités hors population inutilisables qu'on pourrait retrouver dans les unités n'ayant pu être rejointes.

Cette formule est basée sur l'hypothèse que le taux de réponse dans la population visée est similaire à celui de la population initialement échantillonnée. De plus, avec le facteur de correction , on exclut du nombre total des non-répondants les unités inutilisables non identifiées, c'est-à-dire les unités non répondantes qui n'auraient pas pu conduire à des entrevues. Une autre approche aurait été d'estimer le nombre de non-répondants inéligibles (utilisables et inutilisables) et de les soustraire du dénominateur de la formule 3. Cette approche nécessite toutefois plus d'hypothèses et d'estimations que celle que nous proposons, ce qui peut entraîner des erreurs plus importantes dans le calcul des taux. L'approche que nous proposons nous semble plus robuste sous cet angle.

Voici les étapes du calcul de ce taux de réponse. Supposons que les résultats de la cueillette se répartissent de la façon suivante:

a) Échantillon total;
b) Hors population au contrôle de l'échantillon;
c) Hors population au "contact";
d) Non-rejoints;
e) Hors-population durant l'entrevue ou à la validation;
f) Refus, incomplets, absents ou autres raisons;
g) Entrevues complétées.

Calculons d'abord le nombre d'unités inutilisables découvertes (rejointes) au sondage. Il s'agit d'additionner les résultats b) et c).

h) Unités inutilisables rejointes: $b + c$

Estimons ensuite le nombre d'unités inutilisables non-rejointes, qui sera le facteur de correction pour le dénominateur.

i) Facteur de correction: $d \times \frac{c}{a - b - d}$

Le taux de réponse est alors donné par:

$$\frac{g + e}{a - h - i}$$

Ce taux de réponse représente une estimation de l'effort global du groupe d'interviewer pour rejoindre le maximum de personnes visées par le sondage à partir d'un échantillon d'unités appartenant à la population visée.

L'exemple ci-dessous illustre les quatres méthodes présentées dans cet article. Il s'agit d'un sondage téléphonique dont l'échantillon a été produit à l'aide du générateur de numéros de téléphone aléatoires C.O.N.T.A.C.T.[1]. Le sondage était réalisé auprès de la population adulte du Québec mais s'adressait à une petite partie de la population. Le répondant était sélectionné aléatoirement à l'aide d'une grille de sélection de Kish. Pour être éligible à un des questionnaires, le répondant devait d'abord compléter un questionnaire de dépistage d'une durée moyenne de 5 minutes.

EXEMPLE:

Voici les résultats administratifs de la cueillette.

a)	Échantillon total (numéros de téléphone)	21 786
b)	Non rejoints après cinq (5) appels	1 807
c)	Numéros inexistants ou discontinués	3 653
d)	Numéros non résidentiels	2 233
e)	Ligne en trouble	209
f)	Entrevues complétées inéligibles	8 915
g)	Hors strate	167
h)	Langue étrangère ou incapacité physique	379
i)	Absence prolongée	468
j)	Entrevues incomplètes	104
k)	Refus du ménage	1 200
l)	Refus de la personne sélectionnée	538
m)	Entrevues complétées éligibles	2 113

Formule 1: $\frac{\text{Nombre d'entrevues complétées éligibles (m)}}{\text{Échantillon total (a)}} = 9{,}7\%$

Formule 2: $\frac{\text{Nombre d'entrevues complétées éligibles (m)}}{\text{Nombre de contacts (m + l + j)}} = 76{,}7\%$

1. C.O.N.T.A.C.T.: Composition par Ordinateur de Numéros de Téléphone Aléatoires Contrôlée par Territoire. Un produit de SOM inc.

Formule 3: $\frac{\text{Nombre d'entrevues complétées éligibles (m)}}{\text{Échantillon total - nombre d'unités hors popul. (a - (c + d + f + g))}} = 32,0\%$

Formule 4:

$$\frac{\text{Nombre d'entrevues complétées éligibles + Nombre d'entrevues complétées inéligibles (m + f)}}{\text{Échantillon total - Nombre d'unités inutilisables - Facteur de correction } (a - (c + d + g) - \frac{(b + e) \times (c + d)}{a - g - b - e})} = 72,8\%$$

Il s'agit peut-être d'un cas extrême où le taux d'incidence est très faible, mais on voit clairement que les trois premières méthodes ne tiennent pas compte de toutes les unités inéligibles dans l'échantillon de départ. Aucune d'entre elles ne peut rendre compte adéquatement de la qualité de la cueillette vraiment accomplie. Par ailleurs, le facteur de correction de numéros inutilisables permet d'ajuster le dénominateur afin que celui-ci reflète mieux la banque de répondants potentiels. Dans ce cas, il corrige le taux de réponse obtenu pour un échantillon "RDD" et le rend comparable à celui d'un échantillon qui aurait été obtenu par le bottin et où l'on aurait également utilisé la formule proposée pour calculer le taux de réponse.

4. CONCLUSION

Ce bref survol de la problématique du calcul des taux de réponse ainsi que la présentation d'une nouvelle approche montre qu'il est primordial que la présentation de taux de réponse soit accompagnée de tous les résultats de la cueillette ainsi que des motivations qui nous poussent à calculer le taux d'une telle façon plutôt que d'une autre.

De plus, il est nécessaire que les entreprises et organismes impliqués dans le milieu du sondage se dotent de normes établies et reconnues de tous afin de maintenir une image de qualité dans leurs services. Cette normalisation n'aboutira qu'à la suite de milliers d'heures de discussion et nous espérons que ce document réussira à en animer quelques-unes.

BIBLIOGRAPHIE

Wiseman F. et McDonald P., Toward the Development of Industry Standards for Response and Non response Rates. Report No. 80-101, Marketing Science Institute, Cambridge, Mass. 1980.

Kviz, F. J., Toward a Standard Definition of Response Rate. Public Opinion Quarterly (Summer 1977) pp 265-67.

ETUDE EXPÉRIMENTALE SUR L'AMÉLIORATION DU TAUX DE RÉPONSE DANS LES ENQUÊTES POSTALES AUPRÈS DES MÉDECINS

LAMBERT Jean, LEGAULT Claudine, MAHEUX Brigitte

Département de médecine sociale et préventive
Université de Montréal

Introduction

Lors d'enquêtes postales auprès de populations de médecins, il est de plus en plus reconnu qu'un taux de réponse acceptable s'avère difficile à obtenir de nos jours[2]. Cependant il est également reconnu que, chez ces mêmes gens, les non-répondants diffèrent peu ou pas des répondants et que le biais éventuel dans l'estimation de différents paramètres peut être faible[2,7]. Il faut toutefois souligner que ce faible biais a été observé lors d'enquêtes d'intérêt général sur la pratique médicale. Un sujet plus personnel touchant les attitudes, les opinions ou les comportements de ces mêmes médecins pourrait entraîner un biais non négligeable. Dès lors, il s'avérerait essentiel d'utiliser les moyens pour augmenter le taux de réponse afin de réduire ce biais.

Dans le cadre d'une enquête postale portant sur un sujet personnel, soit les attitudes professionnelles d'une population de médecins, nous avons jugé essentiel de mettre en oeuvre les moyens appropriés afin d'obtenir un bon taux de réponse. Outre les moyens habituels, à savoir un nombre suffisant de rappels et un questionnaire simple et attrayant, nous avons expérimenté l'ajout d'un post-scriptum écrit à la main au bas d'une lettre de rappel et remerciant le répondeur. Nous avons également expérimenté un envoi plus personnalisé, à savoir un affranchissement à l'aide de timbres plutôt qu'avec le permis postal sur les enveloppes de départ et de retour et une adresse dactylographiée plutôt que l'utilisation d'étiquettes autocollantes sur l'enveloppe de départ.

L'utilisation d'un post-scriptum, peu étudiée dans la littérature et, à notre connaissance, jamais étudiée chez des médecins, semble pro-

duire en moyenne une amélioration de 5% sur le taux de réponse [4,6,13,16]. L'affranchissement à l'aide de timbres sur l'enveloppe de retour (à l'envoi initial ou lors de rappels), à notre connaissance jamais étudié chez des médecins non plus, produit en moyenne une amélioration de 4% sur le taux de réponse [3,5,8-11,15,17,20,21]. Si l'affranchissement à l'aide de timbres est utilisé sur l'enveloppe de départ, cette amélioration diminue en moyenne à 2% [3,10,12,15]. Quant à l'utilisation d'une adresse dactylographiée (ou à la main) plutôt que d'une étiquette autocollante, trop peu d'études permettent de juger de sa valeur [1,15]. D'autres moyens incitateurs, tel l'ajout d'un crayon dans l'envoi, se sont également montrés efficaces [18].

Méthodologie

La présente expérimentation s'est effectuée dans le cadre d'une enquête postale menée en 1983-1984 auprès de médecins québécois. Cette enquête avait pour but d'étudier les attitudes professionnelles des médecins en fonction de certaines caractéristiques, tels les aspects sociodémographiques, de formation ou de pratique médicales. La population-cible était constituée des médecins omnipraticiens et spécialistes, francophones, résidant au Québec et ayant un contact avec des patients. A l'aide des fichiers administratifs de la Corporation professionnelle des médecins du Québec et de la Fédération des médecins omnipraticiens du Québec, deux échantillons aléatoires ont été recueillis, l'un stratifié selon le sexe et l'enseignement médical (oui, non) pour les spécialistes, l'autre stratifié selon le sexe, la formation médicale (médecine familiale ou non) et le lieu de pratique (C.L.S.C. ou non) pour les omnipraticiens.

Lors de l'envoi initial du questionnaire, tous les spécialistes (n = 379) et tous les omnipraticiens (n = 729) reçurent la même documentation : une lettre de couverture, le questionnaire et une enveloppe de retour préaffranchie mécaniquement, le tout dans une enveloppe également affranchie mécaniquement et adressée à l'aide d'une étiquette autocollante.

Lors de la première relance, les non-répondants à l'envoi initial reçurent de nouveau la même documentation présentée de la même façon; cependant ils furent randomisés en deux groupes afin qu'environ la moitié des sujets reçoivent la lettre de rappel avec le post-scriptum suivant : "Merci, docteur X, pour votre collaboration" (PS^+). L'autre moitié reçut la lettre de rappel sans post-scriptum (PS^-). Lors de la deuxième relance, les non-répondants à la première relance reçurent encore une fois la même documentation, mais ils furent randomisés en deux groupes afin qu'environ la moitié des sujets reçoivent l'envoi personnalisé (PER^+). L'autre moitié reçut un envoi présenté de la même façon que précédemment (PER^-). Il y eut finalement une troisième relance où tous les non-répondants à l'envoi précédent reçurent une dernière fois la même documentation dans un envoi personnalisé (PER^+) et avec l'ajout d'un post-scriptum (PS^+).

A partir des fichiers administratifs, deux autres variables furent recueillies, à savoir l'année d'obtention de la licence médicale pour tous les médecins et le type de spécialité (médecine ou chirurgie) pour les spécialistes. A l'aide d'un modèle de régression logistique multiple, il fut possible d'étudier dans quelle mesure ces variables et les variables de stratification influençaient la probabilité de retourner le questionnaire, et ce, après chaque envoi.

Résultats

On présente au tableau 1 les résultats descriptifs sommaires. On remarque d'abord que les omnipraticiens ont répondu davantage à l'enquête que les spécialistes, et ce, après chaque envoi. En ce qui concerne l'évolution du taux de réponse cumulé, on a observé une augmentation d'environ 14% de l'envoi initial à la première relance et de la première à la deuxième relance, et ce, autant chez les omnipraticiens que chez les spécialistes. Cependant, cette augmentation a diminué à 5.1% et 8.4% entre la deuxième et la dernière relance chez ces deux groupes respectivement, décrivant un effet plateau éventuel.

En ce qui concerne la première expérimentation, le post-scriptum a augmenté les taux de réponse de 7.6% et 10.6% chez les omnipraticiens et les spécialistes respectivement. Il s'agit d'augmentations importantes si l'on considère que, sans l'ajout d'un post-scriptum, les taux de réponse spécifiques furent de 25.3% et 16.3% respectivement pour ces deux groupes, comparativement à 32,9% et 26,9% sans l'ajout d'un post-scriptum. Cela représente des augmentations relatives de 30.0% chez les omnipraticiens et de 65.0% chez les spécialistes par rapport aux taux de réponse obtenus sans post-scriptum.

En ce qui concerne la deuxième expérimentation, l'envoi personnalisé a augmenté les taux de réponse de 12.9% et 14.6% chez les omnipraticiens et les spécialistes respectivement. Il s'agit également d'augmentations importantes. En effet, dans l'envoi non personnalisé, les taux de réponse spécifiques étaient de 33.1% chez les omnipraticiens et de 16.3% chez les spécialistes. Il s'agit d'augmentations relatives de 39.0% et 89.6% respectivement par rapport aux taux obtenus dans l'envoi non personnalisé.

Après chaque envoi où les effectifs le permettaient (i.e., l'envoi initial et les deux premières relances), l'ajustement de modèles de régression logistique multiple permit d'identifier, en plus des variables d'expérimentation, les caractéristiques des médecins pouvant influencer les taux de réponse. Sont présentés aux tableaux 2 et 3 les caractéristiques des médecins et les variables d'expérimentation ayant une contribution significative ($p < 0.05$) sur le taux de réponse après l'envoi initial et les deux premières relances. Les résultats indiquent que les omnipraticiens travaillant dans un centre local de services communautaires, ayant une formation en médecine familiale et étant de sexe féminin avaient une probabilité plus élevée de retourner le questionnaire lors de l'envoi initial ou lors de la première relance (tableau 2). Parmi les variables d'expérimentation, seul l'envoi personnalisé a contribué significativement à l'amélioration du taux de réponse chez les omnipraticiens.

En ce qui concerne les spécialistes, le fait d'être impliqués dans l'enseignement aux étudiants en médecine contribuait à augmenter la probabilité de retourner le questionnaire lors de l'envoi initial (tableau 3). A la première relance, une interaction fut observée entre le sexe et l'année d'obtention de la licence médicale. En effet, les femmes ayant obtenu leur licence vers les années soixante ont moins répondu que les hommes. Cependant cette différence a disparu pour les médecins ayant obtenu leur licence vers les années soixante-dix et a réapparu, mais en favorisant les femmes, pour les médecins ayant obtenu leur licence vers les années quatre-vingt. Tout au long de cette période, le taux de réponse chez les hommes fut relativement stable. Finalement, chez les spécialistes hommes et femmes, les deux variables d'expérimentation ont contribué significativement à l'amélioration du taux de réponse.

Discussion

Les taux de réponse obtenus après trois rappels se comparent très favorablement à l'ensemble des taux de réponse obtenus dans diverses enquêtes postales auprès de médecins et compilées dans deux études [7,19].

Les expérimentations des deux techniques pour améliorer les taux de réponse ont produit de meilleurs résultats que ce que l'on retrouve en général dans les enquêtes qui ont testé les mêmes techniques sur des populations générales ou spécifiques et portant sur divers sujets [3-6,8-13,15-17,20,21]. En effet, l'ajout d'un post-scriptum a amélioré les taux de réponse de 7.6% et 10.6% chez les omnipraticiens et les spécialistes respectivement, en regard de 5% en moyenne dans la littérature [4,6,13,16]. Quant à l'envoi personnalisé (timbres et adresse dactylographiée), la comparaison est plus difficile étant donné que, dans la présente étude, le type d'affranchissement et l'adresse dactylographiée ont été étudiés simultanément. Cependant, on pourrait croire que les améliorations de 12.9% et 14.6% sur les taux de réponse des omnipraticiens et des spécialistes respectivement, en regard de 4% dans la littérature concernant l'affranchissement à l'aide de timbres,

pourraient être principalement dues à l'affranchissement même [3,5,8-11,15,17,20,21]. Toutefois, l'utilisation d'une adresse dactylographiée, sans résultats probants dans la littérature, pourrait s'avérer plus efficace en agissant en interaction avec l'affranchissement et étant destinée à une population (médecins) habituée à recevoir du courrier affranchi avec un permis postal et adressé avec des étiquettes autocollantes. Finalement, on remarque que les augmentations relatives ont été plus importantes chez les spécialistes : tout en représentant des effets "régression à la moyenne", elles dénotent également l'importance d'un post-scriptum ou d'un envoi personnalisé lorsque les taux de réponse spécifiques sont faibles.

Concernant les caractéristiques des médecins ayant une contribution significative sur le taux de réponse, on remarque que leur influence est également importante : en effet, des différences de taux de réponse de près de 30% sont observées dans certains cas. De plus, certaines de ces caractéristiques étant associées aux attitudes professionnelles des médecins [14], les non-répondants seraient donc différents des répondants, ce qui n'appuie pas les résultats précédents [2]. Il est dès lors essentiel de redresser l'échantillon ou de pondérer les résultats. Il faut cependant ajouter, et ceci appuie les résultats précédents [2], que l'apport de nouveaux médecins à l'aide de rappels successifs ne corrige pas de façon appréciable ce biais. Il est donc préférable, dans de telles situations, d'effectuer le nombre de rappels approprié à l'obtention d'une précision (dans les estimations) ou d'une puissance (dans les tests d'hypothèse) acceptables plutôt que de tenter de corriger ce biais par de multiples rappels. Et dans ce cas, la correction du biais s'effectuera par redressement de l'échantillon ou pondération des résultats suite au(x) rappel(s) jugé(s) approprié(s) pour l'obtention de la précision ou puissance voulues. Nous terminons en soulignant qu'il serait cependant hasardeux de vouloir généraliser à d'autres populations ou sujets d'études.

Bibliographie

1. Anderson, J.F., Berdie, D.R., (1975), "Effects on Response Rates of Formal and Informal Questionnaire Follow-up Techniques", J. Applied Psychology, 60, 255-257.

2. Berk, M.L., (1985), "Interviewing Physicians : The Effect of Improved Response Rate", Am. J. Public Health, 75, 1338-1340.

3. Champion, D.J., Sear, A.M., (1969), "Questionnaire Response Rate : A Methodological Analysis", Social Forces, 47, 335-339.

4. Childers, T.L., Pride, W.M., Ferrell, O.C., (1980), "A Reassessment of the Effects of Appeals on Response to Mail Surveys", J. Marketing Research, 17, 365-370.

5. Finn, D.W., (1983), "Response Speeds, Functions and Predictability in Mail Surveys", J. Academy Marketing Science, 11, 61-70.

6. Frazier, G., Bird, K., (1958), "Increasing the Response of a Mail Questionnaire", J. Marketing, 23, 186-187.

7. Gough, H.G., Hall, W.B., (1977), "A Comparison of Physicians Who Did or Did not Respond to a Postal Questionnaire", J. Applied Psychology, 62, 777-780.

8. Gullahorn, J.E., Gullahorn, J.T., (1963), "An Investigation of the Effects of Three Factors on Response to Mail Questionnaires", Public Opinion Quarterly, 27, 294-296.

9. Harris, J.R., Guffey, H.J., Jr., (1978), "Questionnaire Returns : Stamps Versus Business Reply Envelopes Revisited", J. Marketing Research, 15, 290-293.

10. Hensley, W.E., (1974), "Increasing Response Rate by Choice of Postage Stamps", Public Opinion Quarterly, 38, 280-283.

11. Jones, W.H., Linda, G., (1978), "Multiple Criteria Effects in a Mail Survey Experiment", J. Marketing Research, 15, 280-284.

12. Kernan, J.B., (1971), "Are "Bulk-Rate Occupants" Really Unresponsive?", Public Opinion Quarterly, 35, 420-424.

13. Longsworth, D.S., (1953), "Use of a Mail Questionnaire", Am. Sociological Review, 18, 310-313.

14. Maheux, B., Lambert, J., Beaudoin, C., Pineault, R., Legault, C., (1987), "Les médecins formés en médecine familiale pratiquent-ils une médecine différente des autres généralistes?", L'Union médicale du Canada, 16, 243-249.

15. Peterson, R.A., (1975), "An Experimental Investigation of Mail Survey Responses", J. Business Research, 3, 199-210.

16. Pressley, M.M., (1979), "Care Needed When Selecting Response Inducements in Mail Surveys of Commercial Populations", J. Academy Marketing Science, 6, 336-343.

17. Robinson, R.A., Agisim, P., (1951), "Making Mail Surveys More Reliable", J. Marketing, 15, 415-424.

18. Sallis, J.F., Fortmann, S.P., Solomon, D.S., Farquhar, J.W., (1985), "Increasing Returns of Physician Surveys", Am. J. Public Health (Letters to the Editor), 74, 1043.

19. Shosteck, H., Fairweather, W.R., (1979), "Physician Response Rates to Mail and Personal Interview Surveys", Public Opinion Quarterly, 43, 206-217.

20. Watson, J.J., (1965), "Improving the Response Rate in Mail Research",J. Advertizing Research, 5, 45-50.

21. Wiseman, F., (1973), "Factor Interaction Effects in Mail Survey ResponseRates", J. Marketing Research, 10, 330-333.

Tableau 1

Taux de réponse suivant chaque envoi postal chez les omnipraticiens et les spécialistes

ENVOI Expérimentation	OMNIPRATICIENS		SPÉCIALISTES		TOTAL	
	TRS* %	TRC* %	TRS %	TRC %	TRS %	TRC %
Envoi initial						
Total	51.2	51.2	34.6	34.6	45.5	45.5
Première relance						
Total	28.9	65.3	21.4	48.6	25.8	59.6
PS+	32.9		26.9		30.4	
PS-	25.3		16.3		21.6	
Deuxième relance						
Total	39.5	79.0	23.6	60.7	32.6	72.7
PER+	46.0		30.9		39.5	
PER-	33.1		16.3		25.8	
Troisième relance						
Total	24.2	84.1	21.5	69.1	22.8	78.9

* TRS = Taux de réponse spécifique obtenu après chaque envoi

* TRC = Taux de réponse cumulatif

Tableau 2

Probabilités de répondre prédites à partir des modèles de régression logistique multiple chez les omnipraticiens

ENVOI	Variable explicative / Catégorie		PROBABILITÉ DE RÉPONDRE PRÉDITE
ENVOI INITIAL			
	Lieu de pratique	Formation antérieure	
	C.L.S.C.	Traditionnelle	0.57
	C.L.S.C.	Médecine familiale	0.66
	Hors C.L.S.C.	Traditionnelle	0.39
	Hors C.L.S.C.	Médecine familiale	0.49
PREMIÈRE RELANCE			
	Sexe / Lieu de pratique	Formation antérieure	
	F / C.L.S.C.	Traditionnelle	0.44
	F / C.L.S.C.	Médecine familiale	0.70
	F / Hors C.L.S.C.	Traditionnelle	0.20
	F / Hors C.L.S.C.	Médecine familiale	0.43
	H / C.L.S.C.	Traditionnelle	0.24
	H / C.L.S.C.	Médecine familiale	0.49
	H / Hors C.L.S.C.	Traditionnelle	0.09
	H / Hors C.L.S.C.	Médecine familiale	0.23
DEUXIÈME RELANCE			
	Envoi personnalisé		
	NON		0.33
	OUI		0.46

Tableau 3

Probabilités de répondre prédites à partir des modèles de régression logistique multiple chez les spécialistes

ENVOI	Variable explicative / Catégorie		PROBABILITÉ DE RÉPONDRE PRÉDITE
ENVOI INITIAL			
Enseignement			
NON			0.22
OUI			0.43
PREMIÈRE RELANCE			
Sexe	Année de licence	Post-scriptum	
F	1960	Oui	0.13
F	1960	Non	0.07
F	1970	Oui	0.25
F	1970	Non	0.14
F	1980	Oui	0.42
F	1980	Non	0.27
H	1960	Oui	0.27
H	1960	Non	0.16
H	1970	Oui	0.24
H	1970	Non	0.14
H	1980	Oui	0.23
H	1980	Non	0.12
DEUXIÈME RELANCE			
Envoi personnalisé			
NON			0.16
OUI			0.31

SÉLECTION DU RÉPONDANT LORS DE SONDAGES TÉLÉPHONIQUES

COMMUNICATION PRÉSENTÉE
AU COLLOQUE SUR LES APPLICATIONS
DE LA STATISTIQUE, DANS
LE CADRE DU 55e CONGRÈS DE L'ACFAS

ROSAIRE LANDRY
STATISTICIEN
VILLE DE LAVAL

LE 20 MAI 1987

TABLE DES MATIÈRES

INTRODUCTION .....

1. POSITION DU PROBLÈME .....

2. QUELQUES EXEMPLES DE PROCÉDURES DE SÉLECTION .....

   2.1. La méthode de Kish .....

   2.2. La méthode de Troldahl-Carter .....

   2.3. La méthode des jours anniversaires .....

3. LA PROCÉDURE UTILISÉE POUR NOTRE SONDAGE .....

   3.1. Motivations pour une nouvelle procédure .....

   3.2. Une adaptation de la méthode de Kish .....

   3.3. La construction des grilles .....

   3.4. Validation de la procédure .....

CONCLUSION .....

BIBLIOGRAPHIE .....

Tableau 1: Grilles de sélection proposées par Kish (1949) ...

Tableau 2: Exemple de grille de sélection pour la méthode de Troldahl-Carter (1964) .....

Tableau 3: Grilles de sélection utilisées pour notre enquête .....

INTRODUCTION

Dans le cadre d'un sondage téléphonique auprès des Lavallois, nous avons été confrontés au problème de la sélection des personnes à interviewer dans un échantillon de ménages. C'est une question bien connue des spécialistes des sondages, à laquelle on applique généralement de façon routinière des solutions éprouvées de longue date.

Après avoir examiné différentes méthodes de sélection rapportées dans la littérature ou utilisées par des firmes de sondage québécoises, nous avons choisi de développer une procédure adaptée aux particularités de l'étude en cause.

Le développement et l'application d'une telle procédure comprend deux principaux aspects:

- la construction des grilles de sélection;
- les modalités d'utilisation des grilles par les enquêteurs.

Le présent exposé couvre ces deux aspects dans le cadre de l'application particulière qui nous concerne.

Auparavant, nous situons le problème dans un contexte plus général et nous présentons quelques exemples connus de procédures de sélection.

1. POSITION DU PROBLÈME

La plupart des bases de sondage utilisées lors d'enquêtes auprès de la population ont le ménage comme unité d'échantillonnage. C'est le cas notamment de l'annuaire téléphonique et des listes rues-adresses comme le Lovell. Or, dans la majorité des sondages d'opinion ou de comportement, l'unité d'enquête et d'analyse est l'individu. Il faut alors déduire un échantillon de personnes à partir d'un échantillon de ménages.

Le statisticien a le choix entre plusieurs solutions. Kish (1965) suggère cinq approches, allant de la sélection d'une seule personne par ménage avec une probabilité égale à l'inverse de la taille du ménage jusqu'à l'inclusion de toutes les personnes admissibles du ménage, et il en discute les mérites respectifs du point de vue des coûts, des biais possibles et des effets sur la variance des estimations. Certaines des solutions intermédiaires qu'il propose sont intéressantes et elles mériteraient qu'on s'y attarde. Compte tenu toutefois des possibilités de corrélations positives entre les réponses des membres d'un ménage et des risques d'interférences lors d'entrevues multiples, l'option habituellement retenue consiste à sélectionner un seul répondant par ménage.

Comme on ne connait généralement pas à l'avance la composition du ménage, la sélection du répondant s'effectue lors du premier contact de l'enquêteur avec un membre du ménage. Dans le contexte d'un sondage téléphonique, l'interviewer doit disposer d'une procédure de sélection à la fois simple, efficace et qui respecte les critères d'objectivité de l'échantillonnage probabiliste.

2. QUELQUES EXEMPLES DE PROCÉDURES DE SÉLECTION

2.1. La méthode de Kish

C'est Kish (1949) qui, le premier, a publié une méthode pratique et objective de sélection (1). Selon cette méthode, l'enquêteur établit d'abord une liste des personnes admissibles dans le ménage et les numérote de 1 à p en commençant par les hommes, du plus vieux au plus jeune, suivis des femmes, de la plus âgée à la plus jeune. Une grille pré-imprimée sur la page couverture du questionnaire indique à l'enquêteur le rang de la personne à choisir selon le nombre de personnes admissibles du ménage.

Sous sa forme originale, la méthode de Kish utilise huit grilles réparties aléatoirement entre les questionnaires dans les proportions indiquées au tableau 1.

---

(1) La méthode de Kish et d'autres méthodes étaient cependant utilisées bien avant par des maisons de sondage et des agences de statistiques. Deming (1960) décrit une procédure semblable à celle de Kish, qu'il aurait apprise en 1946 chez Alfred Politz Research Inc. à New York.

Tableau 1: Grilles de sélection proposées par Kish (1949)

Proportion des questionnaires	Grille	Si le nombre de personnes admissibles est:					
		1	2	3	4	5	6 ou plus
		Choisir la personne de rang:					
1/6	A	1	1	1	1	1	1
1/12	B1	1	1	1	1	2	2
1/12	B2	1	1	1	2	2	2
1/6	C	1	1	2	2	3	3
1/6	D	1	2	2	3	4	4
1/12	E1	1	2	3	3	3*	5
1/12	E2	1	2	3	4	5*	5
1/6	F	1	2	3	4	5	6

Cette procédure accorde une chance égale d'être choisi à chaque membre admissible d'un ménage de taille donné, sauf aux personnes de rangs 3 et 5 dans les ménages de taille 5 qui sont légèrement sur-représentées (probabilité de 3/12 au lieu de 1/6). On peut facilement obtenir des probabilités de sélection égales à 1/p pour tous en augmentant le nombre de grilles et en modifiant leur répartition. Pour des raisons pratiques, Kish avait plutôt choisi de réduire le nombre de grilles et de les attribuer aux questionnaires dans des proportions simples.

La procédure comporte une seconde limite puisque les personnes admissibles de rang supérieur à six ont une probabilité nulle d'être choisie. Kish (1965) estimait qu'un adulte sur mille (1) se trouvait dans cette situation aux U.S.A. en 1957. Il

(1) Principalement des jeunes femmes, à cause du mode d'énumération adopté.

s'agit donc d'un facteur négligeable, d'autant plus que la taille des ménages n'a cessé de diminuer depuis.

La méthode de Kish est encore souvent utilisée aujourd'hui, avec ou sans modifications. Lorsqu'il y en a, les changements portent surtout sur le mode d'énumération des membres du ménage (e.g.: les femmes avant les hommes) ou sur le nombre et la répartition des grilles. Une firme de Montréal, par exemple, fournit des grilles applicables à des ménages pouvant compter jusqu'à dix personnes admissibles.

2.2. La méthode de Troldahl-Carter

On a souvent reproché à la procédure de Kish d'être un peu lourde et, surtout, d'être trop "indiscrète". Développée dans un contexte d'entrevue à domicile où l'enquêteur peut établir une relation de confiance avec son interlocuteur et prendre le temps d'obtenir la collaboration souhaitée, elle s'est révélée plus difficile d'application dans les sondages téléphoniques. De plus, la multiplication des sondages et des diverses formes de sollicitation téléphonique ont entraîné une plus grande réticence de la population à répondre aux questions des enquêteurs. Face à ce problème et compte tenu des coûts croissants des entrevues, les spécialistes ont cherché à mettre au point des procédures plus expéditives et requérant moins d'informations de la part de la personne qui répond au téléphone.

Troldahl et Carter (1964) ont proposé une méthode où l'enquêteur pose seulement deux questions (1):

«Combien de personnes de 18 ans et plus, y compris vous-même, habitent chez vous ?»

«Parmi ces personnes, combien y a-t-il d'hommes ?»

La sélection s'opère à l'aide d'une grille qui prend la forme d'un tableau à double entrée imprimé sur la page couverture du questionnaire ou annexé à celui-ci et qui indique la personne à interviewer selon les réponses aux deux questions.

Tableau 2: Exemple de grille de sélection pour la méthode de Troldahl-Carter (1964)

Nombre d'hommes	Nombre de personnes admissibles dans le ménage			
	1	2	3	4 ou plus
0	femme	femme la plus âgée	femme la plus jeune	femme la plus jeune
1	homme	homme	homme	femme la plus âgée
2		homme le plus âgé	homme le plus jeune	homme le plus jeune
3			homme le plus jeune	homme le plus âgé
4				homme le plus âgé

(1) La formulation varie quelque peu selon les auteurs. De plus, la limite d'âge spécifiée dépend de la tranche de population visée par le sondage.

Dans la version originale publiée en 1964, Troldahl et Carter utilisaient seulement quatre grilles, attribuées aux questionnaires dans des proportions égales, et où seuls les hommes et les femmes les plus jeunes ou les plus âgés avaient des chances égales d'être choisis. Dans les ménages comportant plus de deux hommes ou plus de deux femmes admissibles, les hommes et les femmes d'âges intermédiaires se trouvaient exclus. Troldahl et Carter avaient estimé à l'époque, que seulement 2 à 3% de la population adulte des États-Unis étaient dans cette catégorie. D'autres études, citées par Bryant (1975), évaluent toutefois aujourd'hui ce pourcentage à près de 5%.

Plusieurs variantes de la méthode de Troldahl-Carter ont été proposées et testées. Bryant (1975) suggère de modifier la répartition des grilles pour accorder une probabilité de sélection plus grande aux hommes afin de compenser leur sous-représentation chronique dans les sondages. Dans une étude récente, Czaja, Blair et Sebestik (1982), ont obtenu un taux de réponse un peu plus élevé en demandant «le nombre de femmes» plutôt que «le nombre d'hommes» comme deuxième question. Cette différence proviendrait de la réticence des femmes seules ou chefs de ménages à indiquer spontanément à un inconnu qu'il n'y a pas d'homme adulte à la maison.

Les améliorations les plus fréquentes consistent à augmenter le nombre de grilles de manière à accorder une chance égale à tous les membres de ménages de 4 personnes et moins (1) ou, comme Groves et Kahn (1979), à étendre les grilles-tableaux jusqu'à un nombre plus grand de personnes admissibles.

---

(1) C'est l'approche que semblent privilégier les firmes de sondage québécoises.

2.3. La méthode des jours anniversaires

Nous terminons ce tour d'horizon avec une approche complètement différente proposée par O'Rourke et Blair (1983) et qui consiste à demander "la dernière personne qui a fêté son anniversaire de naissance" ou, comme version alternative, "la prochaine personne qui fêtera son anniversaire de naissance". D'après leur étude, cette approche, comparée à celle de Kish, diminue sensiblement les refus qui surviennent pendant le processus de sélection. Évidemment, la distribution des mois de naissance des répondants ne correspond pas à celle de la population en général, car elle se concentre autour du mois de l'enquête. Ce n'est pas là cependant un inconvénient majeur à moins que l'on croit à l'influence des signes du zodiaque sur les comportements ou les opinions des individus par rapport à l'objet de l'enquête !

3. LA PROCÉDURE UTILISÉE POUR NOTRE SONDAGE

3.1. Motivations pour une nouvelle procédure

À l'instar de plusieurs auteurs et collègues, la méthode de Kish, sous sa forme originale, nous paraissait un peu trop lourde pour un sondage téléphonique. Quant à la méthode des dates anniversaires, elle se situe nettement en dehors des pratiques habituelles et, malgré son apparente simplicité, elle pose certaines difficultés d'application (1).

D'autre part, notre enquête s'adressant aux Lavallois âgés de 15 ans et plus, les versions à quatre grilles de la méthode de Troldahl-Carter ne semblaient guère appropriées. Comme nous estimions à environ 5% la proportion des ménages lavallois comptant plus de quatre personnes de 15 ans et plus (2), les versions à six grilles, exactes pour des ménages ayant quatre personnes admissibles ou moins, paraissaient tout juste acceptables.

La méthode de Troldahl-Carter étant tout de même la plus populaire, notre première démarche fut de développer des jeux de grilles exactes à quatre et à cinq personnes admissibles (3). Mais, l'extension du nombre de grilles et de leur étendue rend cette méthode assez peu pratique. Des grilles plus larges augmentent

---

(1) Même s'ils sont favorables à cette approche, O'Rourke et Blair (1983) soulignent quelques problèmes et ils souhaitent que d'autres recherches viennent confirmer leurs résultats.

(2) D'après les résultats du sondage, cette proportion se situerait aux alentours de 4,2%.

(3) Un exemple à six personnes admissibles est donné dans Groves et Kahn (1979).

les risques d'erreurs des enquêteurs et amènent une certaine ambiguïté lorsque celui-ci doit demander à parler, par exemple, "à l'homme le 2ème plus âgé" ou "à la femme la 3ème plus jeune". De plus, la taille des grilles est telle qu'il faut les imprimer sur des feuilles séparées et les assembler ensuite avec les questionnaires dans des proportions adéquates. Quelques pré-tests effectués avec le personnel de notre service nous ont d'ailleurs convaincus qu'il fallait chercher une méthode plus "élégante".

3.2. Une adaptation de la méthode de Kish

L'avantage de la procédure de Kish réside dans la simplicité des grilles de sélection. Par contre, la méthode de Troldahl-Carter exige peu d'information sur la composition du ménage; l'enquêteur pose deux questions faciles et pas trop compromettantes pour son interlocuteur. Nous avons donc cherché à mettre au point une procédure alliant ces deux qualités.

Les grilles de Kish sont simples parce que les renseignements demandés par l'enquêteur lui permettent d'induire un ordre strict sur les membres du ménage. Par conséquent, le problème consiste simplement à trouver un moyen d'induire un tel ordre avec un minimum de renseignements.

En abordant le problème sous cet angle, il nous a semblé que la variable âge, prise seule, était suffisante et que l'enquêteur pouvait laisser à son interlocuteur le soin d'ordonner les membres du ménage suivant ce critère.

En partant de ces deux principes, nous avons élaboré une procédure de sélection où:

1°) l'enquêteur demande d'abord à la personne qui prend l'appel, combien il y a de personnes admissibles dans le ménage (15 ans et plus dans notre cas);

2°) l'enquêteur demande ensuite à parler à la personne qui, en plaçant les membres du ménage en ordre décroissant d'âge (1), arrive au rang x spécifié par la grille de sélection.

Il y a 12 grilles de sélection attribuées aléatoirement aux ménages de l'échantillon dans des proportions telles que chaque membre d'un ménage de taille p ait une chance sur p d'être choisi. Ces grilles sont imprimées avec le numéro de dossier et le numéro de téléphone sur des étiquettes auto-collantes qui sont ensuite apposées sur la page couverture des questionnaires.

Pour faciliter la tâche des enquêteurs, nous avons utilisé les codes mnémoniques A et J à la place des numéros d'ordre pour désigner respectivement la personne la plus âgée et la plus jeune parmi les membres admissibles. Nous avons également simplifié la question posée par l'enquêteur lorsque le répondant désigné par la grille était le plus âgé ou le plus jeune. Il s'agit d'une simplification non négligeable puisque d'après les résultats du sondage, nous estimons à seulement 13,3% la proportion des ménages où l'enquêteur devait demander de classer les personnes en ordre d'âge.

Enfin, en travaillant un peu la disposition des éléments qui composent la procédure (la grille, les questions de l'enquêteur et ses instructions), nous avons pu les réunir sur une page

---

(1) Lors d'un pré-test, l'ordre décroissant a semblé plus naturel aux enquêteurs qu'un ordre croissant.

couverture bilingue, un peu chargée mais bien lisible et facile d'emploi pour les enquêteurs. La page suivante en montre une copie réduite.

3.3. La construction des grilles

En partant du principe que pour donner une chance 1/p à chaque membre d'une famille de taille p, le nombre de grilles doit être divisible par chacun des p_i considérés, Paisley et Parker (1965) concluent que ce nombre doit être au moins égal au plus petit commun multiple (P.P.C.M.) des p_i et ils fournissent une méthode pour calculer les 60 grilles requises pour des ménages de tailles 1 à 5.

Or, ce raisonnement n'est pas tout à fait exact car la probabilité de sélection ne résulte pas uniquement du nombre de grilles, mais à la fois de leur nombre et de leur distribution parmi les ménages de l'échantillon. On aura besoin de 60 grilles si on leur impose une distribution uniforme. Autrement, on peut construire une matrice comprenant beaucoup moins de grilles distinctes.

On peut assez facilement obtenir 12 grilles qui font l'affaire à partir du tableau des grilles de Kish. On peut aussi y arriver en utilisant l'idée du P.P.C.M. et en procédant de la façon suivante: sachant que chacun des rangs de 1 à p_i doit apparaître 10, 12, 15, 20 et 30 fois dans une séquence complète pour des valeurs de p_i égales à 6, 5, 4, 3 et 2, on construit une matrice de 60 par 5 dont la première colonne contient 30 fois le nombre 1 suivi de 30 fois le nombre 2, la deuxième colonne contient 20 fois le nombre 1 suivi de 20 fois le nombre 2 et de 20 fois le nombre 3, et ainsi de suite. En regroupant les grilles identiques, on obtient les 12 grilles listées au tableau 3 avec leurs fréquences d'apparition.

Dossier 2500-02-04 Tél: 999-9999

Nbre de personnes: 2 3 4 5 6+

Choisir : A A 2 2 2

Bonjour (bonsoir),
Good (morning, afternoon, evening)

ici _____ [ ][ ]
my name is (18)(19)

de (of) _____

NOUS FAISONS UN SONDAGE POUR LA VILLE DE LAVAL SUR LES LOISIRS CULTURELS DES LAVALLOIS.
WE ARE CONDUCTING A SURVEY ON BEHALF OF THE CITY OF LAVAL ABOUT CULTURAL LEISURE.

D'abord, pour me permettre de choisir un membre de votre famille, j'aimerais savoir combien de personnes de 15 ans et plus habitent chez-vous; n'oubliez pas de vous compter vous-même si vous avez au moins 15 ans.
To help me select a respondent, could you first tell me how many persons aged 15 or over are living in your home, counting yourself if you are 15 or over.

SI 0 ⟶ REMERCIER ET CESSER L'ENTREVUE (MÉNAGE NON ÉLIGIBLE)

1 ⟶ Êtes-vous cette personne ? Are you this person ?

OUI ⟶ J'aimerais maintenant vous poser quelques questions au sujet des arts et de la culture
YES I would like to ask you a few questions about arts and culture.

NON (NO)

2 OU PLUS ⟶ ENCERCLER LE RANG SUR LA GRILLE ET LE REPORTER

SI RANG J: J'aimerais parler à la plus jeune de ces _____ personnes.
I would like to speak to the youngest of these _____ persons.

SI RANG A: J'aimerais parler à la plus âgée de ces _____ personnes.
I would like to speak to the oldest of these _____ persons.

SI AUTRE RANG: En classant ces _____ personnes en ordre d'âge en allant de la plus âgée à la plus jeune, j'aimerais parler à celle qui arrive au _____ ième rang.
Ranking these _____ persons in decreasing order of age from the oldest to the youngest, I would like to speak to the person who comes in the _____ place.

SI LA PERSONNE SÉLECTIONNÉE EST ABSENTE:

⟶ Pouvez-vous me donner son prénom _____ et me dire quand je pourrai la rejoindre: jour _____ heure _____ AM/PM.
Could you give me this person's first name _____ and tell me when it would be possible to call him/her back: day _____ time _____ AM/PM.

RÉSULTATS PROGRESSIFS DES APPELS

Appel	Jour/mois	Heure	Code
1er			
2e			
3e			
4e			
5e			
6e			
7e			
8e			
9e			
10e			
11e			
12e			
(20-21)			(22-23)

CODES D'APPELS

01: Questionnaire complété
02: Refus avant sélection
03: Refus pendant la sélection
04: Refus de l'individu choisi
05: Questionnaire incomplet
06: Pas de réponse (5 coups)
07: Ligne occupée
08: Numéro discontinué
09: Numéro non résidentiel
10: Incapacité (sourd, malade, ...)
11: Langue étrangère
12: Ménage non éligible
13: Absence prolongée
14: Personne choisie absente
15: Autre (préciser: _____)

Remarques: _____

Tableau 3: Grilles de sélection utilisées pour notre enquête

Proportion des questionnaires	Numéro de grille	Si le nombre de personnes admissibles est:				
		2	3	4	5	6
		Choisir la personne de rang:				
1/6	1	1	1	1	1	1
1/30	2	1	1	1	1	2
1/20	3	1	1	1	2	2
1/12	4	1	1	2	2	2
1/15	5	1	2	2	2	3
1/10	6	1	2	2	3	3
1/10	7	2	2	3	3	4
1/15	8	2	2	3	4	4
1/12	9	2	3	3	4	5
1/20	10	2	3	4	4	5
1/30	11	2	3	4	5	5
1/6	12	2	3	4	5	6

Note: La colonne "une personne admissible" a été enlevée car elle n'ajoute aucune information au tableau.

3.4. Validation de la procédure

Afin de pouvoir effectuer une validation indirecte de la procédure de sélection, nous avions prévu trois codes de refus: refus de la personne qui prend l'appel (avant et pendant le processus de sélection), refus de l'individu choisi. Nous aurions pu ainsi comparer nos résultats avec ceux de O'Rourke et Blair (1983) et, éventuellement, avec d'autres sondages à venir. Malheureusement, la consigne à ce chapitre n'a pas été suivie correctement par le personnel de la firme qui a réalisé les entrevues, de telle sorte que cette distinction devenait inutilisable.

La seule validation que nous pouvons invoquer pour le moment provient des enquêteurs qui, malgré qu'ils aient été habitués à une autre procédure, ont apprécié la nouvelle méthode et s'y sont adaptés facilement. Ils n'ont pas constaté de difficulté particulière parmi les ménages enquêtés.

CONCLUSION

Une étude détaillée de plusieurs procédures de sélection des répondants au sein d'un échantillon de ménages nous a conduit à développer une méthode qui combine les avantages des deux procédures les mieux connues et les plus fréquemment utilisées, celles de Kish et de Troldahl-Carter.

Compte tenu du type de grilles employé et du principe sous-jacent, cette nouvelle procédure peut être considérée comme une adaptation originale de la méthode de Kish. Elle est toutefois plus facile à appliquer lors d'un sondage téléphonique car elle requiert moins d'information de la part de la personne qui répond au téléphone et elle simplifie le travail de l'enquêteur.

Utilisée pour la première fois lors d'un sondage auprès de la population lavalloise âgée de 15 ans et plus, la procédure a été bien accueillie par les enquêteurs de la firme qui a réalisé les entrevues.

À la vue de cette expérience, nous croyons que cette procédure peut être avantageusement utilisée dans d'autres circonstances. Il serait souhaitable toutefois que des études plus approfondies viennent en confirmer la pertinence et en évaluer les avantages par rapport aux autres méthodes couramment employées par les sondeurs.

BIBLIOGRAPHIE

BRYANT, B.E. (1975). *Respondent Selection in a Time of Changing Household Composition*. Journal of Marketing Research, Vol. XII, 129-135.

COCHRAN, W.G. (1963). *Sampling Techniques* (second edition). New York. John Wiley & Sons, Inc.

CZAJA, R., BLAIR, J. et SEBESTIK (1982). *Respondent Selection in a Telephone Survey: A comparison of Three Techniques*. Journal of Marketing Research, Vol. XIX, 381-385.

DEMING, W.E. (1960). *Sample Design in Business Research*. New York. John Wiley & Sons, Inc.

DESABIE, J. (1966). *Théorie et pratique des sondages*. Paris. Dunod.

GROVES, R.M. et KAHN, R.L. (1979). *Survey by Telephone: A National Comparison with Personal Interviews*. New York. Academic Press.

KISH, L. (1949). *A Procedure for Objective Respondent Selection within the Household*. Journal of the American Statistical Association, Vol. 44, 380-387.

KISH, L. (1965). *Survey Sampling*. New York. John Wiley & Sons, Inc.

O'ROURKE, D. et BLAIR, J. (1983). *Improving Random Respondent Selection in Telephone Surveys*. Journal of Marketing Research, Vol. XX, 428-432.

PAISLEY, W.J. et PARKER, E.B. (1965). *A Computer-Generated Sampling Table for Selecting Respondents within Households*. Public Opinion Quarterly, Vol. 29, 431-436.

SATIN, A. et SHASTRY, W. (1983). *L'Échantillonnage: un guide non mathématique*. Ottawa. Statistique Canada.

TROLDAHL, V.C. et CARTER, R.E.JR (1964). *Random Selection of Respondents within Households in Phone Surveys*. Journal of Marketing Research, Vol. 1, 71-76.

ENQUETE SUR LA CONDUITE AUTOMOBILE AVEC LES FACULTES AFFAIBLIES PAR L'ALCOOL - 1986

QUEBEC

Allocution présentée le 19 mai 1987

au colloque de l'ACFAS

à l'Université d'Ottawa

Pierre Paquet
Direction de la statistique
Vice-présidence à la planification et
à la promotion de la sécurité routière
Régie de l'assurance automobile du Québec

Mai 1987

La présente allocution porte sur l'enquête effectuée en 1986 concernant la conduite automobile avec les facultés affaiblies par l'alcool au Québec. Il s'agissait plus particulièrement d'obtenir une évaluation de la proportion des gens qui conduisent avec les facultés affaiblies ainsi que de déterminer certaines caractéristiques de ces derniers pour en dresser le profil, dans le but d'appliquer éventuellement des programmes appropriés de sécurité routière. De plus, l'enquête 1986 a permis de voir l'évolution de la situation par rapport à celle de 1981.

Cette enquête a été effectuée dans le cadre d'une enquête nationale coordonnée par Transports Canada et qui s'étend sur deux ans, 1986 et 1987.

Au Québec, elle fut réalisée par la Direction de la statistique de la Régie de l'assurance automobile, avec la coopération de Transports Canada pour l'échantillonnage, les alcootests et le traitement informatique des données, du ministère des Transports du Québec pour l'équipement routier (cônes, tours d'éclairage, panneaux de signalisation, etc.) et du Solliciteur général pour l'incitation auprès des corps policiers à collaborer à l'enquête. De plus, la Sûreté du Québec a participé à la formation des enquêteurs concernant la Loi C-19, le calibrage et l'utilisation des alcootests.

Une façon d'évaluer l'incidence de la consommation d'alcool dans les accidents de véhicules routiers est de comparer, au cours d'une période donnée, la proportion de conducteurs sur les routes en état d'ébriété avec la proportion de conducteurs en état d'ébriété impliqués dans les

accidents. Dans l'éventualité où il n'y a pas de différence significative entre ces deux éléments, il pourrait être conclu que la contribution de l'alcool dans les accidents de véhicules routiers est négligeable.

Tout en obtenant certaines caractéristiques des conducteurs en état d'ébriété sur les routes du Québec, l'enquête de 1986 permettra de tenter un tel exercice.

METHODOLOGIE

Pour la réalisation de cette enquête, 160 sites furent échantillonés au Québec; 16 sites par région à l'exception de la région de Montréal pour laquelle 32 sites furent choisis. Un site est un emplacement hors-route (station d'essence, terrain de stationnement, etc.) permettant d'accueillir au moins trois véhicules à la fois et sur lequel le travail des enquêteurs devait être effectué en toute sécurité.

L'enquête s'est déroulée au cours de la période du 6 au 30 août, du mercredi au samedi, entre 21 heures et 3 heures du matin. Chaque site fut utilisé une seule fois, durant une période d'une durée de 1 heure 30 minutes.

Pour faire les interviews, trois équipes de sept personnes (1 chef d'équipe et 6 enquêteurs) furent formés. Chaque équipe fut divisée en deux groupes de 3 enquêteurs et chaque groupe travaillait en alternance pour une période de 1 heure 30 minutes. Pour effectuer les interviews, trois enquêteurs travaillaient simultanément sur un site donné. Chaque enquêteur disposait d'un alcootest (type J4, à lecture digitale) pour déterminer la concentration d'alcool dans le sang. L'enquête a été effectuée sur une base totalement volontaire de la part des participants.

N'étant pas autorisé à intercepter des véhicules sur la route, la Régie de l'assurance automobile s'est assurée au préalable de la collaboration des corps policiers pour chacun des sites sur lesquels l'enquête devait se dérouler (Sûreté du Québec pour 12 sites et 42 corps policiers municipaux pour 148 sites). La présence policière était absolument nécessaire, sans laquelle l'opération était annulée.

Au cours de l'enquête, lorsque la concentration d'alcool dans le sang (CAS) d'un conducteur était supérieure à la limite légale (80 mg/100 ml), ce conducteur était fortement incité à prendre une des options suivantes:

- changer de conducteur;
- téléphoner à un autre conducteur;
- stationner et prendre un taxi (aux frais de la Régie);
- taxi - deux chauffeurs (aux frais de la Régie);
- stationner et quitter autrement.

Par conséquent, lors de la préparation de l'enquête il a fallu s'assurer pour chaque site utilisé que le service de taxi deux chauffeurs était disponible.

Pour la réalisation de l'enquête, chaque équipe de 7 personnes disposait d'un véhicule de promenade et de 2 petits camions avec la boîte arrière fermée pour le transport de matériels (panneaux de signalisation, cônes de sécurité, vestes de sécurité, tour d'éclairage, génératrice, essence, etc.).

Lors d'une journée d'enquête, le chef d'équipe devait s'assurer sur les lieux de la disponibilité du site, confirmer la présence des policiers ainsi que la disponibilité du service de taxi deux chauffeurs. De plus, avant chaque nuit d'enquête, tous les alcootests étaient calibrés. Sur chaque site, l'installation s'effectuait environ 3/4 d'heure avant le début des interviews. Pour les interviews, au signal des enquêteurs le policier interceptait le nombre de véhicules demandés et les dirigeait vers les enquêteurs.

RESULTATS

Le nombre de véhicules échantillonnés au Québec au cours de l'enquête 1986 fut de 6 104. De ce nombre 5 214 acceptèrent de donner un échantillon d'haleine, soit 85,4%.

Le sommaire des résultats présenté en annexe montre que de 1981 à 1986 il y a eu une amélioration sensible au Québec concernant la conduite nocturne avec les facultés affaiblies par l'alcool, et cela d'une façon plus significative pour les conducteurs en état d'alcoolémie (CAS[1] > 80 mg/100 ml) en passant de 5,9% en 1981 à 3,6% en 1986.

La ventilation des résultats selon différentes variables permet de constater que même si la conduite avec les facultés affaiblies par l'alcool chez les femmes demeure inférieure à celle des hommes, le pourcentage de conductrices en état d'alcoolémie est passée de 1,1% en 1981 à 2,6% en 1986, alors que chez les conducteurs de sexe masculin il a diminué sensiblement (6,9% et 3,8% en 1981 et 1986 respectivement). Il est à noter qu'au cours des périodes d'observations les conducteurs masculins sont surreprésentés, comptant pour 80,5% des conducteurs sur la route alors que la population des titulaires de permis de conduire au Québec est composée de conducteurs de sexe masculin dans une proportion de 58%.

Quant à l'âge des conducteurs, le groupe des 40-59 ans est celui qui montre la plus forte proportion des conducteurs en état d'alcoolémie en 1986 (5,2%). En 1981, c'était les groupes des 25-39 ans et 60 ans et

---

1. CAS: Concentration d'alcool dans le sang.
Ebriété: Etat dans lequel la CAS > 20 mg d'alcool/100 ml de sang.
Alcoolémie: Etat dans lequel la CAS > 80 mg d'alcool/100 ml de sang.

plus qui dominaient avec 7,4% et 7,5% respectivement. Pour les 16-24 ans, alors qu'ils représentent 19% des titulaires de permis de conduire, on les retrouve dans une proportion de 34% des conducteurs entre 21h00 et 03h00. Leur pourcentage en état d'alcoolémie au cours de cette période en 1986 fut estimé à 2,6%.

Concernant les différentes tranches d'heures, c'est la dernière, soit celle de 01h30-03h00, qui présente la proportion la plus élevée de conducteurs sur la route en état d'ébriété et d'alcoolémie en 1981 et 1986. En 1986, ces proportions furent de 29,6% pour l'état d'ébriété et 7,0% pour l'état d'alcoolémie.

En ce qui regarde le port de la ceinture de sécurité, il ressort que plus la présence d'alcool est élevée, moins la ceinture est portée. En 1986, 2,5% étaient en état d'alcoolémie parmi les conducteurs qui portaient la ceinture de sécurité; ce pourcentage était de 7,0% pour les non-porteurs.

CONCLUSION

Cette enquête a permis de constater qu'il y a eu, de 1981 à 1986, un gain significatif sur les habitudes de conduite avec les facultés affaiblies par l'alcool au Québec. Pratiquement sur tous les plans, des gains ont été réalisés.

En utilisant les résultats obtenus dans cette enquête avec ceux d'une étude sur l'"usage de l'alcool par les conducteurs blessés mortellement dans la circulation, de 1983 à 1984" publiés par Transports Canada en avril 1986, portant sur les dossiers d'accidents de 7 provinces[2] il peut être établi que le risque d'un conducteur en état d'alcoolémie (CAS 80 mg/100 ml) de décéder dans un accident de la route est environ 15 fois celui d'un conducteur dont la CAS est inférieur à 80 mg/100 ml. Cette même valeur relative pour les conducteurs de 16-19 ans est de l'ordre de 75 fois.

Tous les traumatismes causés par les pertes de vie dans les accidents de véhicules routiers ainsi que le capital humain de haute qualité perdu pour la société font en sorte qu'il demeure très important de continuer l'application de mesures préventives et éducationnelles par les différents intervenants en sécurité routière.

---

2. Colombie-Britannique, Alberta, Saskatchewan, Manitoba, Ontario, Nouveau-Brunswick et Ile-du-Prince-Edouard.

ANNEXE

ENQUETE SUR LA CONDUITE AUTOMOBILE AVEC FACULTES AFFAIBLIES PAR L'ALCOOL - 1986

QUEBEC

SOMMAIRE DES RESULTATS

Période d'enquête:		6 au 30 août 1986, du mercredi au samedi, de 21h00 et 03h00
Nombre de sites retenus		160
Nombre de sites sur lesquels l'opération s'est déroulée		153
Nombre de véhicules sélectionnés		6 104
Echantillon d'haleine manquant dû à:		
- Refus de répondre au questionnaire	274	
- Refus de fournir un échantillon d'haleine	308	
- Souffle insuffisant	252	
- Equipement déficient	56	
- Total		890
Nombre de conducteurs retenus pour l'analyse		5 214

PROPORTION DES CONDUCTEURS AVEC LES FACULTES AFFAIBLIES PAR L'ALCOOL

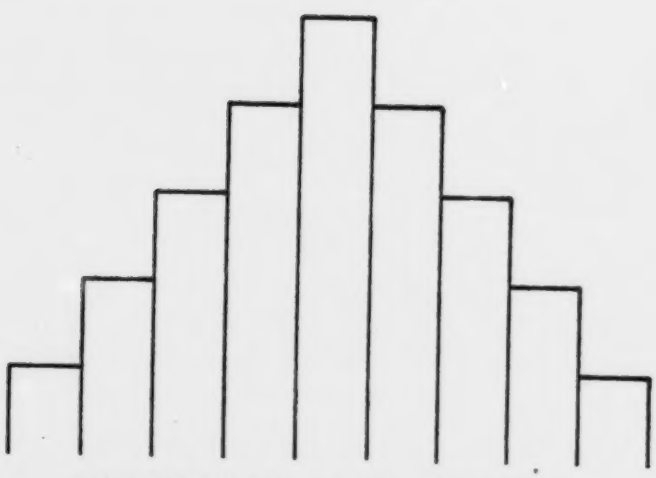
	Etat d'ébriété (CAS[1] > 20mg/100ml)		Etat d'alcoolémie (CAS[1] > 80mg/100ml)	
	1981 %	1986 %	1981 %	1986 %
Ensemble des conducteurs	20,5	17,0	5,9	3,6
Sexe des conducteurs:				
Masculin	21,9	17,8	6,9	3,8
Féminin	14,0	13,1	1,1	2,6
Age des conducteurs				
16-24 ans	21,8	15,1	4,4	2,6
25-39 ans	22,7	19,1	7,4	3,9
40-59 ans	17,0	18,2	5,3	5,2
60 ans et plus	10,1	6,9	7,5	1,7
Heures d'enquête				
21h00 - 22h29	7,9	12,0	2,1	2,1
22h30 - 23h59	23,1	16,4	5,2	3,2
00h00 - 01h29	28,7	20,0	10,3	5,3
01h30 - 03h00	34,8	29,6	10,3	7,0
Nuit d'enquête				
Mercredi	22,5	15,0	6,3	2,1
Jeudi	17,4	14,5	5,0	3,7
Vendredi	21,1	20,6	6,3	5,2
Samedi	24,7	19,0	5,6	3,4
Ceinture de sécurité				
Portée		14,7		2,5
Non portée		23,9		7,0

1. CAS: Concentration d'alcool dans le sang.

Gouvernement du Québec
Ministère du Revenu
Direction générale de la planification
et de la budgétisation

Utilisation de l'échantillonnage avec probabilités proportionnelles à la taille (P.P.S.): exemple d'application sur un fichier administratif

Myriam Proulx
Service des recherches statistiques
Mai 1987



DIRECTION DES ÉTUDES,
RECHERCHES ET STATISTIQUES

DRM-223
84-01

Introduction

Les statisticiens sont souvent amenés à effectuer des enquêtes par sondage pour réaliser des études statistiques. La collecte des données est habituellement une étape de ces enquêtes qui s'avère ardue et qui demande beaucoup de ressources. Les moyens utilisés pour obtenir les données sont généralement la poste, le téléphone ou l'interview de personne à personne. Dans ces situations, les contraintes sont considérables, les coûts sont énormes, le temps est un facteur important et dans plusieurs cas, la non-réponse est un problème.

Cependant, au ministère du Revenu du Québec, comme dans plusieurs autres ministères ou organismes, les fichiers administratifs constituent une source de données extrêmement importante. L'exploitation de ces fichiers à des fins statistiques ou à des fins de vérification fiscale devrait être envisagée régulièrement. C'est une avenue très intéressante car la collecte des données est en grande partie finalisée et la non-réponse n'existe pas. De plus, la capacité de traitement informatique a progressé considérablement depuis une dizaine d'années. La manipulation de fichiers de grande dimension est devenue relativement simple.

Les travaux du Service des recherches statistiques sur ce sujet remontent à 1982, avec un premier rapport de Monsieur Claude Boivin [BOIVIN] proposant l'utilisation de l'échantillonnage avec probabilités proportionnelles à la taille dans les sondages de vérification. La méthode proposée a été appliquée avec succès par Madame Caouette en 1985 [CAOUETTE] et moi-même en 1986 [PROULX], pour l'établissement de la provision pour créances douteuses.

Cette communication montre comment le Service des recherches statistiques utilise un échantillon, avec probabilités proportionnelles à la taille, tiré d'un fichier administratif dans le but d'estimer le total des créances irrécupérables. Une brève description du problème à résoudre constitue la première section et la méthodologie choisie est expliquée à la deuxième section. La troisième section présente l'application de la méthode (plan de sondage, cueillette et traitement des données) et la dernière section est une courte discussion sur les résultats obtenus.

1. Problème à résoudre: estimation du montant total de créances irrécouvrables

Chaque année, le ministère du Revenu doit établir une provision pour créances douteuses pour les comptes à recevoir (CAR) débiteurs. Un CAR débiteur est un montant dû au ministère du Revenu par un contribuable ou une entreprise et ce montant comprend les droits, les intérêts sur droits, les intérêts cycliques, les peines et les chèques retournés. La provision pour créances douteuses consiste en une réserve monétaire créée à la fin d'une année financière pour couvrir les CAR qui seront jugés irrécouvrables à la suite d'une enquête.

Etant donné qu'il y a des milliers de comptes à recevoir, il est préférable de procéder par échantillonnage car il est insensé de penser à faire, chaque année, l'analyse de milliers de comptes en terme de temps et de coûts. L'échantillonnage permet d'obtenir une estimation du taux de provision pour créances douteuses c'est-à-dire le quotient entre le montant total de provision pour créances douteuses et le montant total de comptes à recevoir. Il est important de bien évaluer la variation annuelle de la provision car celle-ci est portée aux dépenses dans l'état financier du ministère.

2. Méthode utilisée: échantillonnage avec probabilités proportionnelles à la taille

Le problème à résoudre consiste à estimer un quotient de deux masses monétaires, quotient compris entre 0 et 1. Une créance douteuse est considérée d'autant plus importante à détecter que le compte à recevoir est élevé. La distribution des comptes à recevoir révèle qu'il y a une grande quantité de petits comptes et peu de comptes élevés.

Deux méthodes ont été envisagées: l'échantillonnage aléatoire stratifié et l'échantillonnage avec probabilités proportionnelles à la taille (pps). Ces deux méthodes permettent d'obtenir l'estimateur désiré. L'échantillonnage avec probabilités proportionnelles à la taille a été préféré à l'échantillonnage aléatoire stratifié pour les raisons pratiques suivantes:

1. élimination de la définition des bornes de stratification;

2. élimination de la mise-à-jour de ces bornes afin de conserver une stratification optimale chaque année;

3. il n'est pas nécessaire de trier les comptes à recevoir selon la variable de stratification;

4. il n'est pas nécessaire de calculer les tailles d'échantillon de chaque strate selon l'allocation de Neyman par exemple.

Généralement, l'échantillonnage avec probabilités proportionnelles à la taille est utilisé lorsque la mesure de grandeur, c'est-à-dire la valeur d'une variable auxiliaire X, est reliée à la caractéristique Y étudiée.

Dans le cas présent, une unité, avec une grande valeur pour la variable Y, contribue davantage au total de la population qu'une petite unité. Il est naturel d'espérer obtenir, dans un tel contexte, de meilleurs estimateurs avec un plan d'échantillonnage où la sélection des unités favorise l'inclusion des grandes unités dans l'échantillon à l'encontre des petites. De plus, la base de sondage était bornée inférieurement et supérieurement. De cette façon, la probabilité d'obtenir de très petites unités était nulle et de même pour les très grandes unités.

Supposons l'estimation du total Y d'une population de N unités selon un échantillonnage avec probabilités proportionnelles à la taille.

On tire n échantillons de taille 1 (S) dans une population U de N unités; on connaît les tailles des unités $M = (M_1, \ldots, M_N)$ où

$$M_0 = \sum_{k=1}^{N} M_k$$

$$\forall k \in U,\ R_k = 1 \text{ si } k \in S \qquad \text{avec } P(R_k = 1) = \frac{M_k}{M_0}$$

$$= 0 \text{ sinon}$$

a) Pour chaque échantillon de taille 1, l'estimateur du total est le suivant:

$$\hat{Y}_{(1)} = \sum_{k=1}^{N} \frac{R_k y_k}{M_k} \cdot M_0 = \frac{y_{(1)}}{M_{(1)}} \cdot M_0$$

$\dot{Y}_{(1)}$ est un estimateur non-biaisé du total car $E[R_k] = \frac{M_k}{M_0} \quad k \in S$

b) Avec n échantillons indépendants de taille 1, on obtient n estimations non-biaisées de Y. Ces estimations peuvent être regroupées au moyen d'un seul estimateur:

$$\dot{Y} = \frac{1}{n} \sum_{l=1}^{n} \dot{Y}_{l(1)} = \frac{1}{n} \sum_{l=1}^{n} \sum_{k=1}^{N} \frac{R_{kl}\, y_{kl}}{M_{kl}} \times M_0$$

C'est aussi un estimateur non-biaisé du total Y.

L'estimateur de la variance de l'estimateur du total se définit comme suit:

$$v(\dot{Y}) = \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(\frac{y_k}{M_k} \cdot M_0 - \dot{Y} \right)^2$$

[MURTHY, p. 255] [COCHRAN, p.252 et ss.]

Le principal avantage d'un tirage avec remise vis-à-vis un tirage sans remise réside dans le calcul de l'estimateur de la variance sans biais. Les formules du calcul de la variance et de l'estimateur de la variance, principalement, sont très simplifiées:

<u>avec remise</u>

$$v(\dot{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{\pi_i} - \dot{Y} \right)^2$$

<u>sans remise</u>

$$v(\dot{Y}) = \sum_{i=1}^{n} \frac{1 - \pi_i}{\pi_i^2} \cdot y_i^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \cdot y_i y_j$$

π_i : probabilité d'inclusion de l'unité i

π_{ij} : probabilité jointe d'inclusion des unités i et j

[BREWER and HANIF, pp.5-8]

Un autre avantage à ne pas négliger se rapporte à la procédure de sélection. On n'a pas à se préoccuper d'éliminer de la base de sondage les unités déjà sélectionnées.

3. Application de la méthode: plan de sondage, cueillette et traitement des données

Soit Y le total à estimer et M_k, le montant du kième compte à recevoir représentant la variable auxiliaire. Celle-ci étant mesurée en dollars,on utilise le sondage selon l'unité monétaire, bien connu dans le milieu de la vérification comptable. Avec cette méthode, la probabilité de sélection d'un compte à recevoir est proportionnelle au montant du compte à recevoir; ainsi elle augmente les chances d'obtenir des gros comptes à recevoir à l'encontre des petits, ce fait étant considéré lors de l'estimation.

Administrativement, les comptes à recevoir sont traités selon les deux grands bureaux régionaux auquels ils sont assignés: Montréal et Québec. On divise donc la population en deux parties, c'est-à-dire selon les deux bureaux régionaux. La taille de l'échantillon déterminée pour l'ensemble de la population est répartie proportionnellement au montant total des comptes à recevoir à l'intérieur de chaque bureau.

Pour fins d'échantillonnage, on utilise un fichier administratif soit le fichier informatisé des CAR. Ce fichier est mis à jour à chaque mois et il comprend tous les comptes à recevoir antérieurs à cette date.

A titre d'exemple, j'utiliserai les données et les résultats de l'étude faite en 1984-1985 portant sur le système intégré des taxes [CAOUETTE].

Base de sondage
(comptes à recevoir compris entre 2 000 $ et 100 000 $ au 31 octobre 1984 pour la catégorie "taxes"*)

Bureau régional	Nombre CAR	Montant (000 $)	Répartition du montant total par bureau (%)
Montréal	3 712	45 116	59
Québec	2 800	31 892	41
TOTAL	6 512	77 008	100

* Faillites exclues

La taille de l'échantillon est déterminée en fonction de la précision absolue désirée, soit ± 5,00%, pour l'estimateur du taux de provision global, et du risque α associé à l'échantillonnage, fixé à 5% donc pour un niveau de confiance de 95%.[COCHRAN, p.77]

Une fois la taille de l'échantillon déterminée, les unités sont prélevées informatiquement. La méthode de sélection utilisée est celle de Lahiri qui correspond à un tirage avec remise et ne nécessite pas l'accumulation du montant de chaque compte à recevoir, c'est-à-dire cumuler les comptes un à un et posséder tous les sous-totaux.

La procédure de sélection de Lahiri consiste à associer à chaque unité de la base de sondage (CAR) un nombre d'unités monétaires (nombre de dollars) égal au montant maximum d'un compte à recevoir. De ces unités monétaires, on considère effectives, c'est-à-dire déterminantes pour la sélection, les unités qui corrrespondent au montant de ce compte à recevoir. Le nombre total d'unités monétaires effectives coïncide ainsi au montant total de tous les comptes à recevoir. Cette procédure équivaut à un tirage avec remise. On utilise le même processus pour chaque strate de la population, c'est-à-dire pour chaque bureau régional.

Posons N = nombre d'unités de la base de sondage

M = montant maximum des comptes à recevoir

M_k = montant du compte à recevoir de l'unité k, k=1,...,N

À chaque unité k, k=1,...,N, on associe M nombres parmi lesquels seulement M_k seront effectifs. Il y a donc N x M nombres associés à la liste des N comptes à recevoir.

L'algorithme de sélection de Lahiri est le suivant:

1. Sélectionner aléatoirement un nombre k entre 1 et N

2. Sélectionner aléatoirement un nombre $ entre 1 et M

3. Si $ $\leq M_k$, l'unité k est choisie
   Si $ $> M_k$, l'unité k est rejetée

et on répète l'algorithme jusqu'à l'obtention du nombre d'unités fixé.

On peut montrer qu'avec cette technique, la probabilité inconditionnelle de sélection de l'unité k est M_k/M_0 où M_0 est le montant total de tous les comptes à recevoir de la base de sondage. [MURTHY, pp. 202-205]

Exemple d'application de l'algorithme de Lahiri

1. k = 2

2. $ = 65 250

3. M_2 = 72 400 d'où $ $\leq M_2$ car 65 250 $\leq$ 72 400
   alors l'unité 2 est choisie.

Le tableau suivant présente les résultats de la sélection:

Tailles d'échantillon des comptes à recevoir

Bureau régional	Nombre de comptes à recevoir
Montréal	225
Québec	152
TOTAL	377

Lorsque l'on a obtenu l'échantillon, les comptes à recevoir sélectionnés sont évalués un à un à l'aide d'une étude qui établit la provision, c'est-à-dire le montant de dollars jugés irrécupérables pour chaque compte. Cette étude est appelée analyse de créance. Les causes sont nombreuses: le contribuable ou le propriétaire de l'entreprise est décédé, il n'est plus au Québec (il peut être à Beyrouth!), il est incarcéré dans un pénitencier, il est insolvable, etc. Le problème de la non-réponse n'existe pas. Une fois l'analyse de créance terminée, le total de dollars irrécupérables est estimé par les formules déterminées à la section 2.

4. Discussion sur les résultats obtenus

Voici les résultats obtenus:

Estimation pour créances douteuses
(comptes à recevoir entre 2 000 $ et 100 000 $)

Bureau régional	Provision Montant (000 $)	Erreur-type Montant (000 $)	Coefficient de variation (%)
Montréal	28 876	1 649	5,71
Québec	20 335	1 197	5,89
TOTAL	49 211	2 038	4,14

Taux de provision et précision
(comptes à recevoir entre 2 000 $ et 100 000 $)

Bureau régional	Taux	Précision*
Montréal	0,6400	± 0,0716
Québec	0,6376	± 0,0736
GLOBAL	0,6390	± 0,0519

* pour un niveau de confiance de 95%

Alors 49,211 M$ sur 77,008 M$ de CAR représentent des créances douteuses. Ce sont des comptes qui sont vieux d'au moins 6 mois.

Les taux de provision pour Montréal et Québec sont considérés égaux étant donné que les intervalles de confiance se chevauchent. De plus, le taux global s'établit à 63,90% (i.e. 63,90% du montant total des CAR est irrécouvrable) avec une marge d'erreur de 5,19% pour un niveau de confiance de 95%. L'échantillonnage répond bien aux objectifs fixés étant donné la précision atteinte.

Conclusion

Les comptes à recevoir élevés, jugés irrécouvrables, contribuent davantage à la provision. Ainsi, la sélection des comptes élevés, peu nombreux, devait être favorisée étant donné la corrélation entre la provision et le montant du compte à recevoir.

Alors les résultats obtenus pour l'exemple présenté dans ce rapport sont très satisfaisants. L'échantillonnage avec probabilités proportionnelles à la taille-tirage avec remise, de par tous ses avantages cités précédemment, s'avère une méthode efficace. Elle est une méthode simple à utiliser sur des fichiers administratifs. Au ministère du Revenu, la même méthode est appliquée pour les corporations, les entreprises faisant des remises de déduction à la source et les particuliers. Cependant, il faut être conscient que l'exploitation des fichiers administratifs suppose l'existence d'un environnement informatique adéquat.

RÉFERENCES

BOIVIN, C. (1982), Sondages de vérification: Échantillonnage dans une population à faible taux d'erreur, rapport #SRS028, Revenu Québec, Québec.

BREWER, K. R. W.
HANIF, M., (1983), Sampling with unequal probabilities, Sringer-Verlag, New-York Inc., New-York.

CAOUETTE, H., (1985), Provision pour créances irrécouvrables des comptes à recevoir de catégorie "entreprises": DECOR, MIDAS, SINTAX, méthodologie d'échantillonnage, Revenu Québec, Sainte-Foy.

COCHRAN, W. G., (1977), Sampling Techniques, third edition, John Wiley, New-York.

LESLIE, D. A.
TEITLEBAUM, A. D.
ANDERSON, R. J., (1980), Dollar-Unit Sampling, Pitman Publishing Limited, London.

MURTHY, M. N., (1967), Sampling Theory and Methods, Calcutta, Statistical Publishing Society.

PROULX, M., (1986), Provision pour créances douteuses des comptes à recevoir de catégorie "individus": COREC, méthodologie d'échantillonnage, Revenu Québec, Sainte-Foy.

L'UTILISATION DE L'ESTIMATEUR PAR LE QUOTIENT POUR MESURER LES TENDANCES.

I. Schiopu-Kratina
K.P. Srinath

DMEE, Statistique Canada

0. Introduction

Dans ce document nous examinons l'utilisation des estimateurs type quotient pour mesurer les tendances. Plusieurs estimateurs sont proposés (voir la section 2) et leurs variances sont comparées (voir la section 3). Des conditions pour que la variance des estimateurs type quotient pour mesurer les tendances (formule (4) - (5)) soit inférieure à la variance de l'estimateur simple pour mesurer les tendances se trouvent également dans la section 3. Ces conditions sont ensuite exprimées en termes de coefficients de corrélation de la variable à estimer et la variable auxilière, à l'occasion t et à deux occasions successives (voir la formule (9)). Si le coefficient de corrélation de la variable à mesurer et la variable auxilière calculé à l'occasion t est notamment plus grand que le coefficient de corrélation de ces variables à deux occasions consécutives, l'estimateur par le quotient est préférable à l'estimateur simple. Ce résultat est présenté à la section 5 à l'aide d'un exemple.

Pour faciliter les calculs et l'interprétation, nous avons fait des hypothèses supplémentaires de symmétrie et de stabilité des estimateurs dans le temps (voir la section 4).

Nous démontrons aussi que la condition obligeant la variance de l'estimateur type quotient pour les tendances à être inférieure à la variance de l'estimateur simple est plus restrictive que la condition correspondante pour l'estimation des niveaux. L'exemple 1 a été conçu pour illustrer cette différence.

1. L'estimateur par le quotient

Soit y la caractéristique à mesurer et x une caractéristique connue. On suppose que le coefficient de corrélation $\rho_{x,y}$ est strictement positif.

Si $\hat{Y}$ est un estimateur de y et $\hat{X}$ est estimateur de x, on définit l'estimateur par le quotient $\hat{Y}_Q$ par la formule:

$$\hat{Y}_Q = \frac{\hat{Y}}{\hat{X}} X \tag{1}$$

Dans la formule (1), X est égale à l'espérance de $\hat{X}$, c.à.d. $X = E\hat{X}$.

Un résultat classique nous dit que la variance de $\hat{Y}_Q$ est plus petite que la variance de $\hat{Y}$ si et seulement si:

$$\rho_{\hat{x},\hat{y}} > \frac{1}{2} \frac{c.v.\hat{X}}{c.v.\hat{Y}} = k \qquad (2)$$

Dans la formule (2), $c.v.\hat{X}$ est le coefficient de variation de $\hat{X}$ et $c.v.\hat{Y}$ est le coefficient de variation de $\hat{Y}$. Le résultat classique mentionné ci-dessus peut être écrit de façon abrégée:

$$\text{Var } \hat{Y}_Q < \text{Var } \hat{Y} \qquad \text{ssi } \rho_{\hat{X},\hat{Y}} > k$$

Pour la preuve, voir, par exemple [1] p. 157.

2. Estimateurs du changement (tendance)

Soit maintenant $\hat{Y}(t)$ un estimateur de la caractéristique y à l'occasion t et soit $Y(t) = E\hat{Y}(t)$, $t = 0, 1, 2, \ldots$. On cherche à estimer le changement $\Delta Y = Y(t) - Y(t-1)$, entre deux occasions successives, $t \geq 1$. Naturellement, le premier estimateur à considérer est défini par la formule:

$$\Delta\hat{Y} = \hat{Y}(t) - \hat{Y}(t-1) \qquad (3)$$

Nous avons considéré deux autres estimateurs, $\Delta\hat{Y}_Q$ et $\Delta^{Q}\hat{Y}$ définis comme suit:

$$\Delta\hat{Y}_Q = \hat{Y}_Q(t) - \hat{Y}_Q(t-1) \qquad (4)$$

$$\Delta^{Q}\hat{Y} = \frac{\hat{Y}(t) - \hat{Y}(t-1)}{\hat{X}(t) - \hat{X}(t-1)} |X(t) - X(t-1)| \qquad (5)$$

où $\hat{Y}_Q(t)$ provient de la formule (1).

3. Comparaison des variances

Pour faciliter les calculs, nous faisons l'hypothèse suivante:

$$\frac{Y(t)}{X(t)} = \frac{Y(t-1)}{X(t-1)} = R \qquad (6)$$

En appliquant la formule de Taylor à la fonction à deux variables $f(x,y) = \frac{y}{x}$ au point $(X(t), Y(t))$ et en négligeant les dérivées de degré supérieur à 1, on obtient (voir aussi (2.41) et [1]) :

$$\hat{Y}_Q(t) \triangleq Y(t) - R\,\hat{X}(t) + \hat{Y}(t) .$$

Puisque $Y(t)$ est un paramètre pour tout $t \geq 0$,

$$\text{Var } \Delta\hat{Y}_Q \triangleq \text{Var } \Delta^Q\hat{Y} \triangleq \text{Var } \{\hat{Y}(t) - \hat{Y}(t-1) - R[\hat{X}(t) - \hat{X}(t-1)]\} . \tag{7}$$

Donc, comme à la formule (2):

$\text{Var } \Delta\hat{Y}_Q < \text{Var } \Delta\hat{Y}$ ssi:

$$\rho_{\Delta\hat{Y},\Delta\hat{X}} > \frac{1}{2}\,\frac{\text{c.v.}\Delta\hat{X}}{\text{c.v.}\Delta\hat{Y}} . \tag{8}$$

Cette formule est équivalente à la formule suivante:

$$\text{Cov } [\hat{Y}(t) - \hat{Y}(t-1),\ \hat{X}(t) - \hat{X}(t-1)] > R/2\ \text{Var } [\hat{X}(t) - \hat{X}(t-1)] . \tag{8'}$$

La formule (8') se prête mieux à des manipulations algébriques, vu que la covariance est une fonction bilinéaire.

4. Comparaison des formules (2) et (8)

On se pose maintenant la question suivante: si, pour toute $t \geq 0$, $\text{Var } \hat{Y}_Q(t) < \text{Var } \hat{Y}(t)$, est-ce-que $\text{Var } [\hat{Y}_Q(t) - \hat{Y}_Q(t-1)] < \text{Var } [\hat{Y}(t) - \hat{Y}(t-1)]$?

La réponse semble être négative. Ceci s'interprête comme suit. Quoique l'estimateur par le quotient $\hat{Y}_Q(t)$ ait une variance plus petite que l'estimateur "simple" $\hat{Y}(t)$ pour la mesure du niveau à tout instant t, il se peut que l'estimateur simple soit préférable à l'estimateur par le quotient pour mesurer les tendances.

Pour simplifier les concepts, nous faisons maintenant des hypothèses de symmétrie et de stabilité dans le temps. Supposons donc que, pour $t \geq 1$, $\text{Var } \hat{X}(t) = \text{Var } \hat{X}(t-1)$, $\text{Var } \hat{Y}(t) = \text{Var } \hat{Y}(t-1)$, $\rho_{\hat{X}(t),\hat{Y}(t)} = \rho_{\hat{X},\hat{Y}} > 0$, $\rho_{\hat{X}(t),\hat{Y}(t-1)} = \rho_{\hat{X}(t-1),\hat{Y}(t)} = \rho_{\hat{X},\hat{Y}}(t,t-1)$.

Notons $\rho_{\hat{X}(t),\hat{X}(t-1)} = \rho_{\hat{X}}(t,t-1) > 0$. Les hypothèses faites sur les coefficients de corrélation et la condition que la matrice de covariance soit positivement définie ne fournissent que la condition connue $\rho_{\hat{X},\hat{Y}}(t,t-1) \leq 1$.

Par contre, il est raisonable de supposer que:

$$\rho_{\hat{X},\hat{Y}}(t,t-1) < \min \{\rho_{\hat{X},\hat{Y}}, \rho_{\hat{X}}(t,t-1)\} .$$

De la formule (8'), on en déduit:

Var $\Delta\hat{Y}_Q$ < Var $\Delta\hat{Y}$ ssi:

$$\rho_{\hat{X},\hat{Y}} > k + \rho_{\hat{X},\hat{Y}}(t,t-1) - k\, \rho_{\hat{X}}(t,t-1) . \qquad (9)$$

Une comparaison avec la formule (2) indique que l'inégalité (9) est plus restrictive que (2) ssi $\rho_{\hat{X},\hat{Y}}(t,t-1) - k\, \rho_{\hat{X}}(t,t-1) > 0$.

En effet, l'exemple suivant présente une telle situation.

Exemple 1

Soit $k = 0.5$, $\rho_{\hat{X},\hat{Y}} = 0.51$, $\rho_{\hat{X}}(t,t-1) = 0.95$, $\rho_{\hat{X},\hat{Y}}(t,t-1) = 0.5$. Puisque $0.51 > 0.5$, Var $\hat{Y}_Q(t)$ < Var $\hat{Y}(t)$ à la suite de (2). Par contre, le terme à droite de l'inégalité (9) est égal à 0.525 et donc (9) n'est pas satisfaite. Il s'en suit que Var $\Delta\hat{Y}_Q$ > Var $\Delta\hat{Y}$.

5. Une condition suffisante pour que (9) soit vrai

Il n'est pas facile de comparer (9) et (2). Il est clair, par exemple, que si la valeur de $\rho_{\hat{X},\hat{Y}}$ est très grande par rapport à la valeur de $\rho_{\hat{X}}(t,t-1)$, il est préférable d'utiliser $\Delta\hat{Y}_Q$ à la place de $\Delta\hat{Y}$. Pourtant, ceci n'est pas le cas en pratique. La condition à suivre permet d'explorer des situations plus réalistes. Vu que $\rho_{\hat{X}}(t,t-1) \geq \rho_{\hat{X},\hat{Y}}(t,t-1)$, il suffit de vérifier:

$$\rho_{\hat{X},\hat{Y}} > k - k\, \rho_{\hat{X},\hat{Y}}(t,t-1) + \rho_{\hat{X},\hat{Y}}(t,t-1) \qquad (10)$$

pour conclure que (9) est satisfaite.

Soit $\alpha = \rho_{\hat{X},\hat{Y}}(t,t-1) / \rho_{\hat{X},\hat{Y}} < 1$

Alors (10) devient:

$$\rho_{\hat{X},\hat{Y}} > \frac{k}{1 + k\alpha + \alpha} . \qquad (10')$$

Pour un k fixe, $1 + k\alpha - \alpha$ est une fonction décroissante de α (de (2) il s'en suit que $k < 1$), donc $\frac{k}{1 + k\alpha - \alpha}$ est une fonction croissante et son maximum de 1 est atteint à $\alpha = 1$.

Exemple 2.

Soit $k = 0.5$, $\alpha = 0.6$. Le quotient à droite de l'inégalité (10') est égal à 0.714....

Si $\rho_{\hat{X},\hat{Y}} = 0.72$ et $\rho_{\hat{X},\hat{Y}}(t,t-1) = 0.43$, alors (10'), et donc (9), est satisfaite. Par la suite, Var $\Delta^{Q}\hat{Y}$ < Var $\Delta \hat{Y}$.

6. Conclusion

Pour décider si l'emploi de l'estimateur (4) est justifié pour mesurer les tendances, il s'en suit qu'il est nécessaire de comparer les coefficients de corrélation $\rho_{\hat{X},\hat{Y}}$, $\rho_{\hat{X},\hat{Y}}(t,t-1)$ et $\rho_{X}(t,t-1)$ à l'aide de la formule (9). Étant donné que la différence entre les formules (9) et (10) est minimale, on peut réduire cette comparaison à la vérification de la formule (10) (ou (10')).

On suggère de calculer les valeurs de $\rho_{\hat{X},\hat{Y}}$ et $\rho_{\hat{X},\hat{Y}}(t,t-1)$ en se basant sur des données observées. Par la suite, si la formule (10') est satisfaite, on suggère d'utiliser l'estimateur définit par la formule (4). Sinon, il est préférable d'utiliser l'estimateur donné par la formule (3) pour mesurer les tendances.

Bibliographie

1. Cochran, W.G., Sampling Techniques, New York, Wiley 1977

Statistics Canada / Statistique Canada
Ottawa, Canada

L'étalonnage et l'interpolation dans les instituts de statistique

par

Pierre A. CHOLETTE

direction pédagogique: Robert POIRIER

Présentation faite au
Colloque sur les applications
de la statistique

organisé dans le cadre du
51e congrès de l'A.C.F.A.S.
à l'université d'Ottawa
du 19 au 22 mai 1987

Division de la Recherche et
de l'analyse des chroniques
Statistique Canada
édifice R.H. Coats 13e étage, "J"
OTTAWA (Canada) K1A 0T6

(613) 991-1600

Canada

BIBLIOGRAPHIE SUR L'ÉTALONNAGE ET L'INTERPOLATION

- Baldwin, A. (1980) "Extensions of Quadratic Minimization Methods", Gross National Product Division, Statistics Canada, Research Paper.

- Bassie, B.L. (1939), "Interpolation Formulae for the Adjustment of Index Numbers," Proceedings of the Annual Meetings of the American Statistical Association.

- Boot, J.C.G., Feibes, W., Lisman, J.H.C. (1967), Further Methods of Derivation of Quarterly Figures from Annual Data", Applied Statistics, Vol. 16, no. 1, pp. 65-75

- Box, G.E.P., Jenkins, G.M. (1970), Time series Analysis, Forecasting and Control, Holden-Day.

- Cholette, P.A. (1978), "Comparaison et évaluation de quelques méthodes d'ajustement de séries infra-annuelles aux repères annuels", Séries chronologiques, Statistique Canada, document de recherche 78-03-001B.

- Cholette, P.A. (1979), "Adjustment Methods of Sub-Annual Series to Yearly Benchmarks", Proceedings of the Computer Science and Statistics, 12th Annual Symposium on the Interface, J.F. Gentleman Editor, University of Waterloo, pp. 358-366.

- Cholette, P.A. (1982), "Prior Information and ARIMA Forecasting", *Journal of Forecasting*, Vol. 1, pp. 375-383.

- Cholette, P.A. (1984), "L'ajustement des séries infra-annuelles aux repères annuels", Techniques d'enquêtes, Statistique Canada, Vol. 10, no. 1, pp. 39-53.

- Cholette, P.A. (1985), "L'estimation des séries infra-annuelles à partir de taux de croissance et de jalons annuels", Séries chronologiques, Statistique Canada, Cahier de travail RASC-85-018B

- Cholette, P.A. (1985), "Linking Heterogeneous Segments of Seasonal Time Series", *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 178-183.

- Cholette, P.A., Piche, R. (1986), "Views on Benchmarking and Interpolation", Time Series, Statistics Canada, Working Paper no. TRSA-86-018E

- Chow, G.C., Lin, An-loh (1971), "Best linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series", Review of Economics and Statistics, Vol. 53, No. 4, pp. 372-375.

- Dagum, E.B. (1977), "Comparison of Various Interpolation

Procedures for Benchmarking Economic Time Series", Séries chronologiques, Statistique Canada, document de recherche 77-05-006E.

- Denton, F.T. (1971), "Adjustment on Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization", Journal of the American Statistical Association, Vol. 66, No. 333, pp. 99-102.

- Fernandez, R.B. (1981), "A Methodological Note on the Estimation of Time Series", Review of Economic and Statistics, Vol. 63, pp. 471-476.

- Friedman, M. (1962), "The Interpolation of Time Series by Related Series", Journal of the American Statistical Association, Vol. 57, No. 300, pp. 729-757.

- Glejser, H. (1966), "Une methode d'évaluation des données mensuelles à partir d'indices trimestriels ou annuels", Cahier Economiques de Bruxelles, No. 19, 1er trimestre, pp. 45-64.

- Ginsburg, V.A. (1973), "A Further Note on the Derivation of Quarterly Figures Consistent with Annual Data", Applied Statistics, Vol. 22, No. 1, pp. 368-374.

- Laniel, N.(1986), "Adjustment of Economic Production Sub-Annual Series to Corresponding Annual Series", Business Survey Methods, Statistics Canada, Working Paper no. BSMD-86-006E

- Lisman, J.H.C., Sandee, J. (1964), "Derivation of Quarterly Figures from Annual Data", Applied Statistics, Vol. 13, No. 2, pp. 87-90.

- Litterman, R.B. (1983), "A Random Walk, Markov Model for the Distribution of Time Series", Journal of Business and Economic Statistics, Vol. 1 No. 2, pp., 169-173.

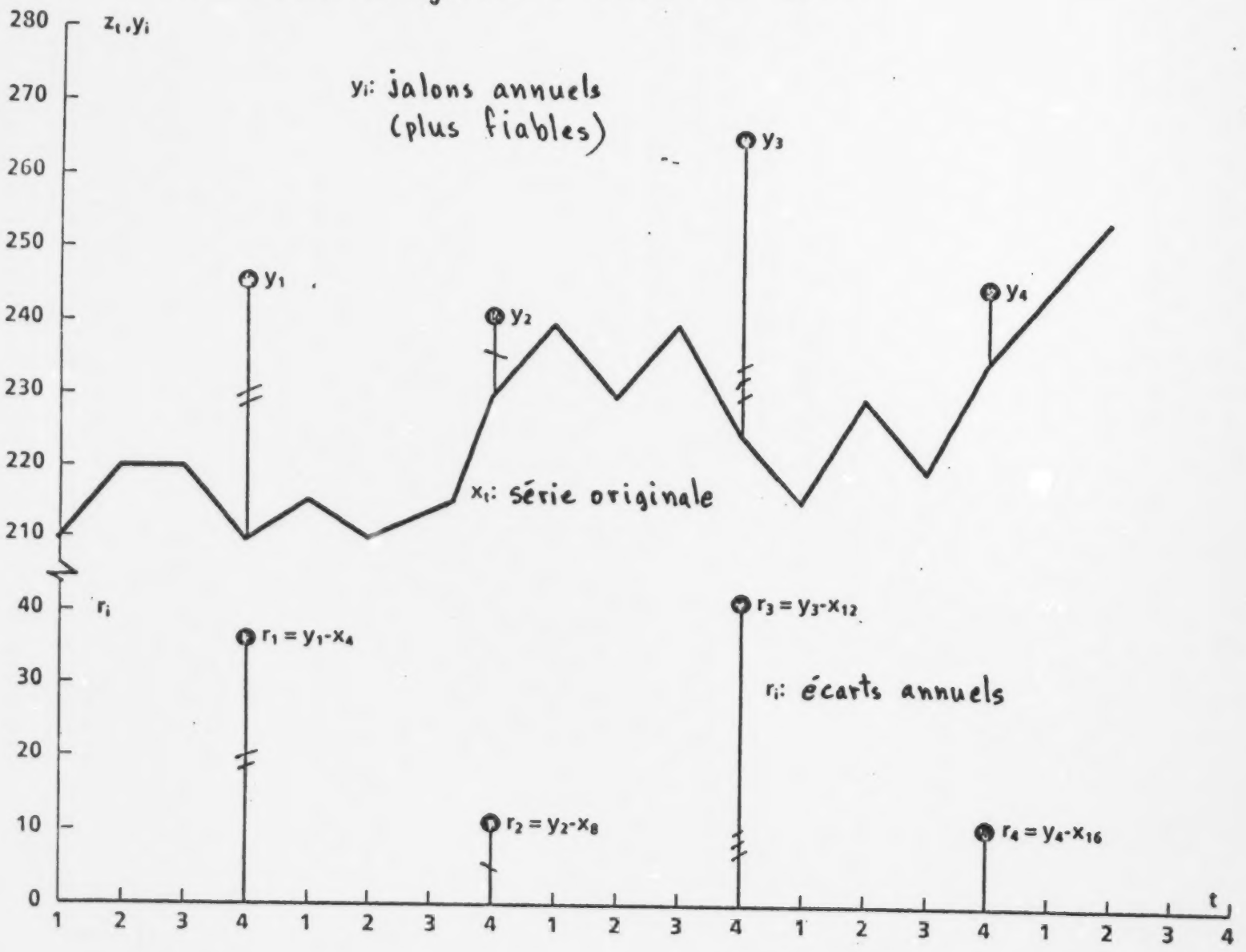
- Monsour, N.J., Trager, M.L. (1979), "Revision and Benchmarking of Business Time Series", Proceedings of the Business and Economic Statistics Section, American Statistical Association.

- Sanz, R. (1981), "Metodos de desagregacion temporal de series economicas", Servicio de Estudios, Banco de Espana.

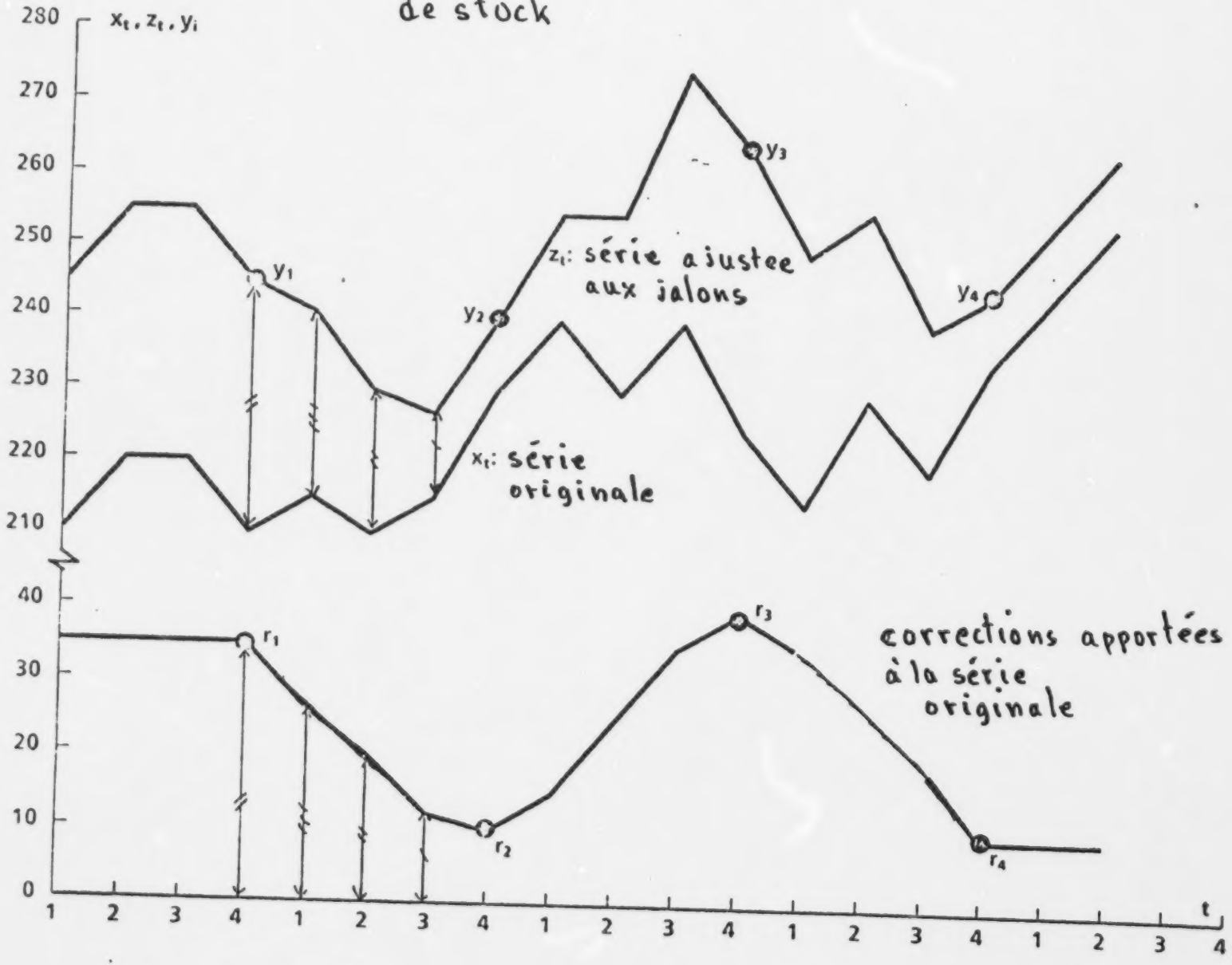
- Somermeyer, W.H., Jansen, R., Lauter, A.S. (1976), "Estimating Quarterly Values from Annually Known Variables in Quarterly Relationships", Journal of the American Statistical Association, Vol. 71, No. 355, pp. 588-595.

- Wilcox, J.A. (1983), "Disaggregating Data Using Related Series", Journal of Business and Economic Statistics, Vol. 1. No. 3, pp. 187-191.

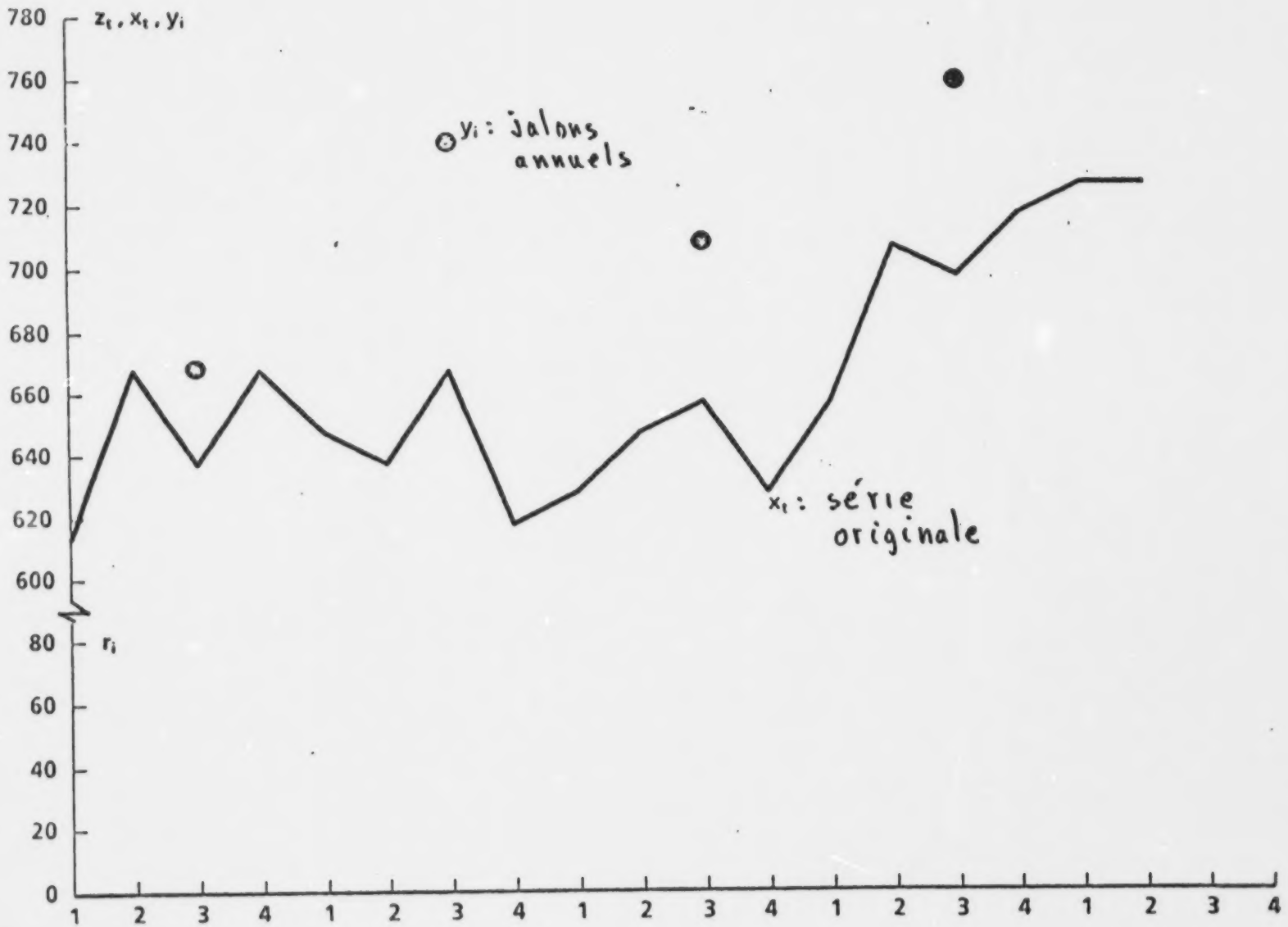
Situation d'étalonnage d'une série de stock



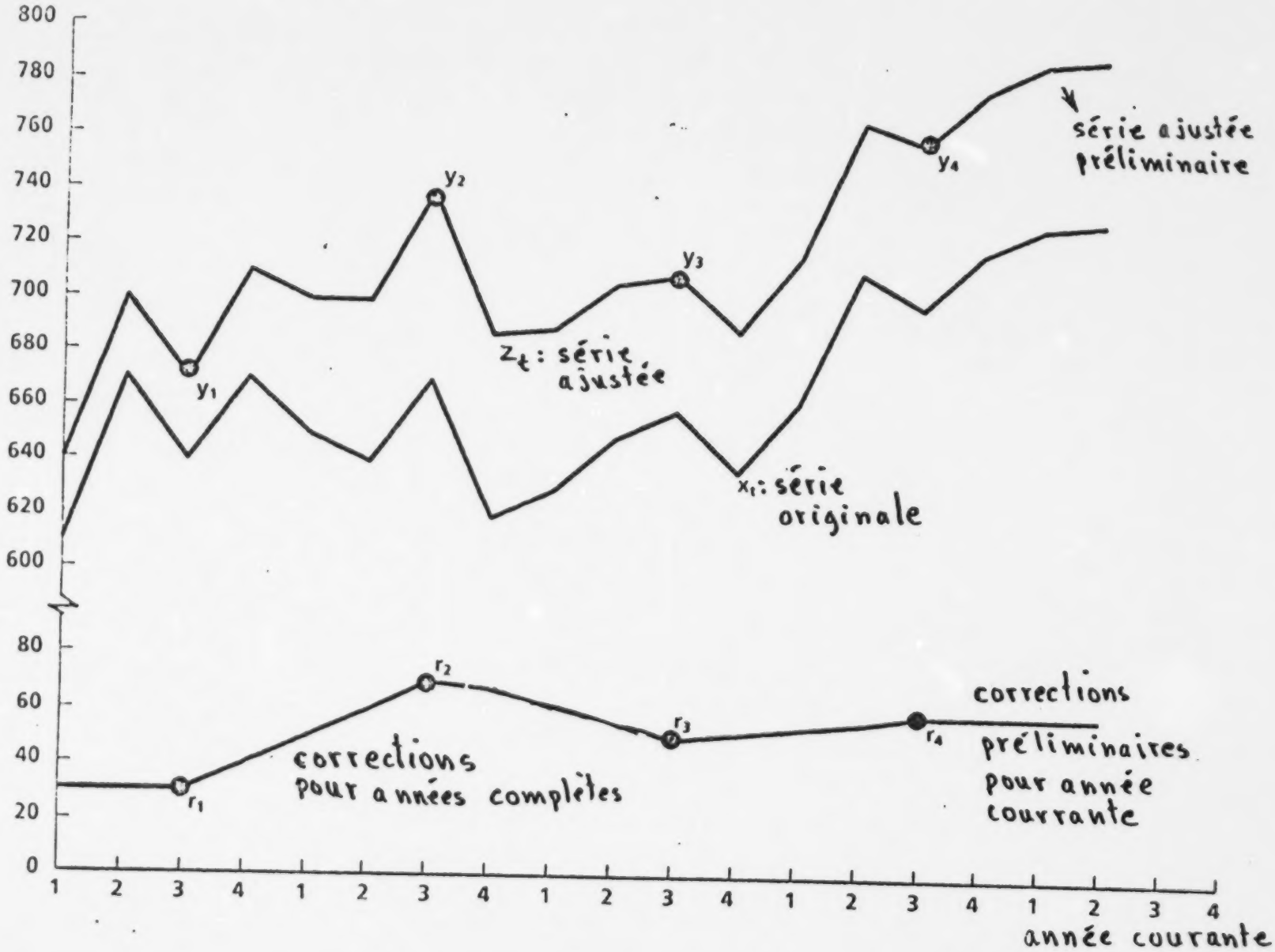
Solution au problème d'étalonnage d'une série de stock



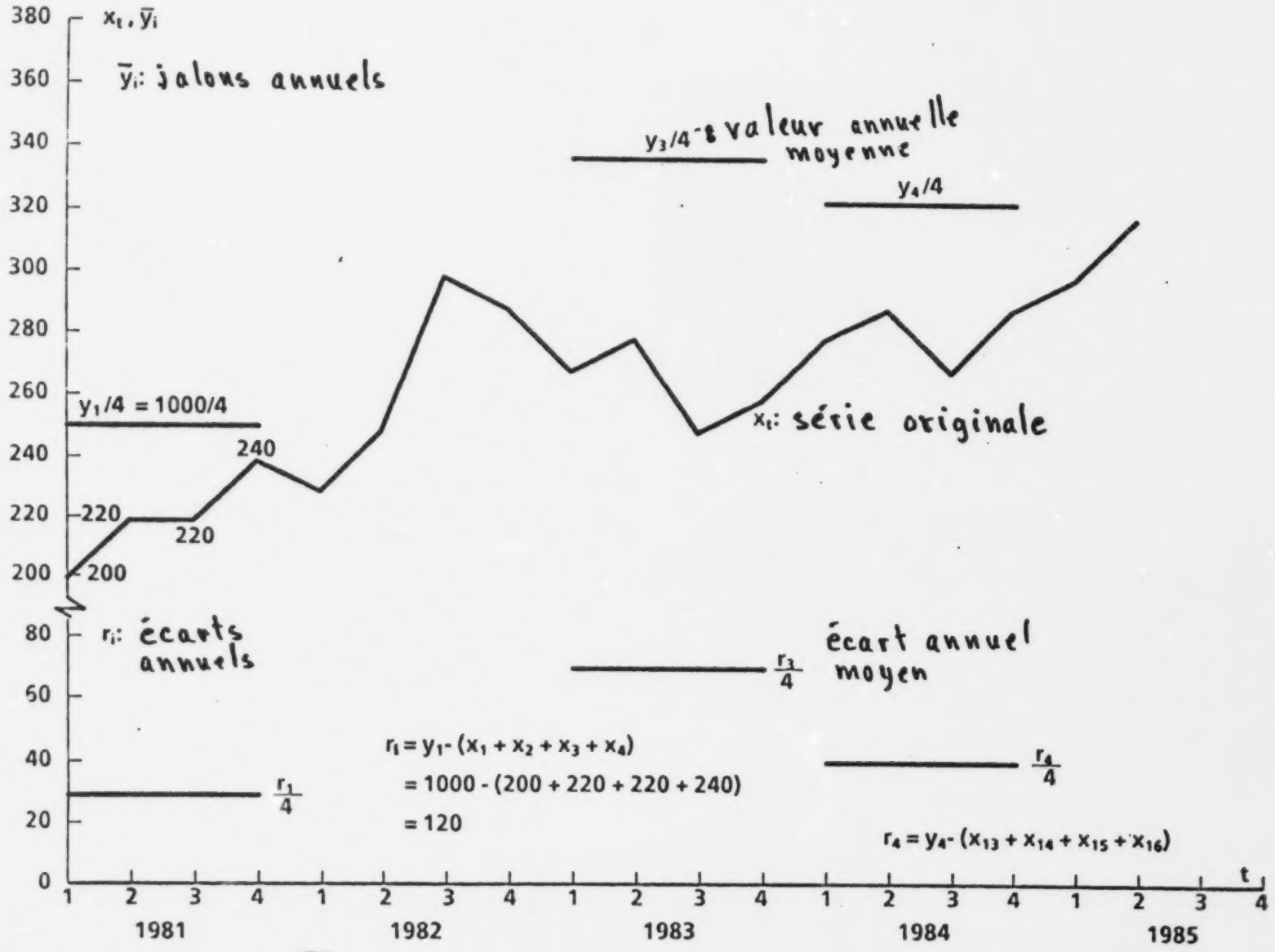
Exercise sur l'étalonnage des series de stock



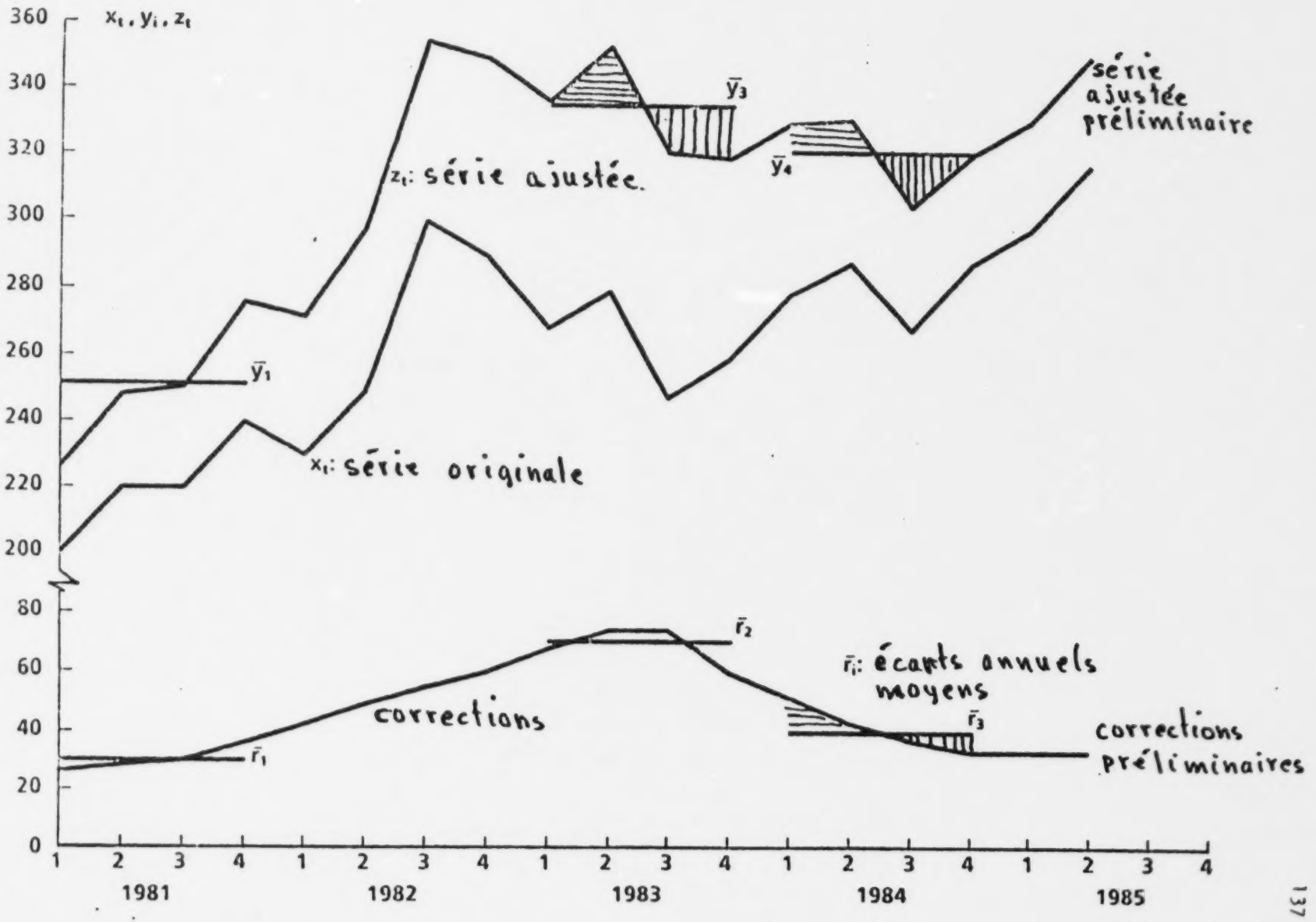
Réponse avec commentaires



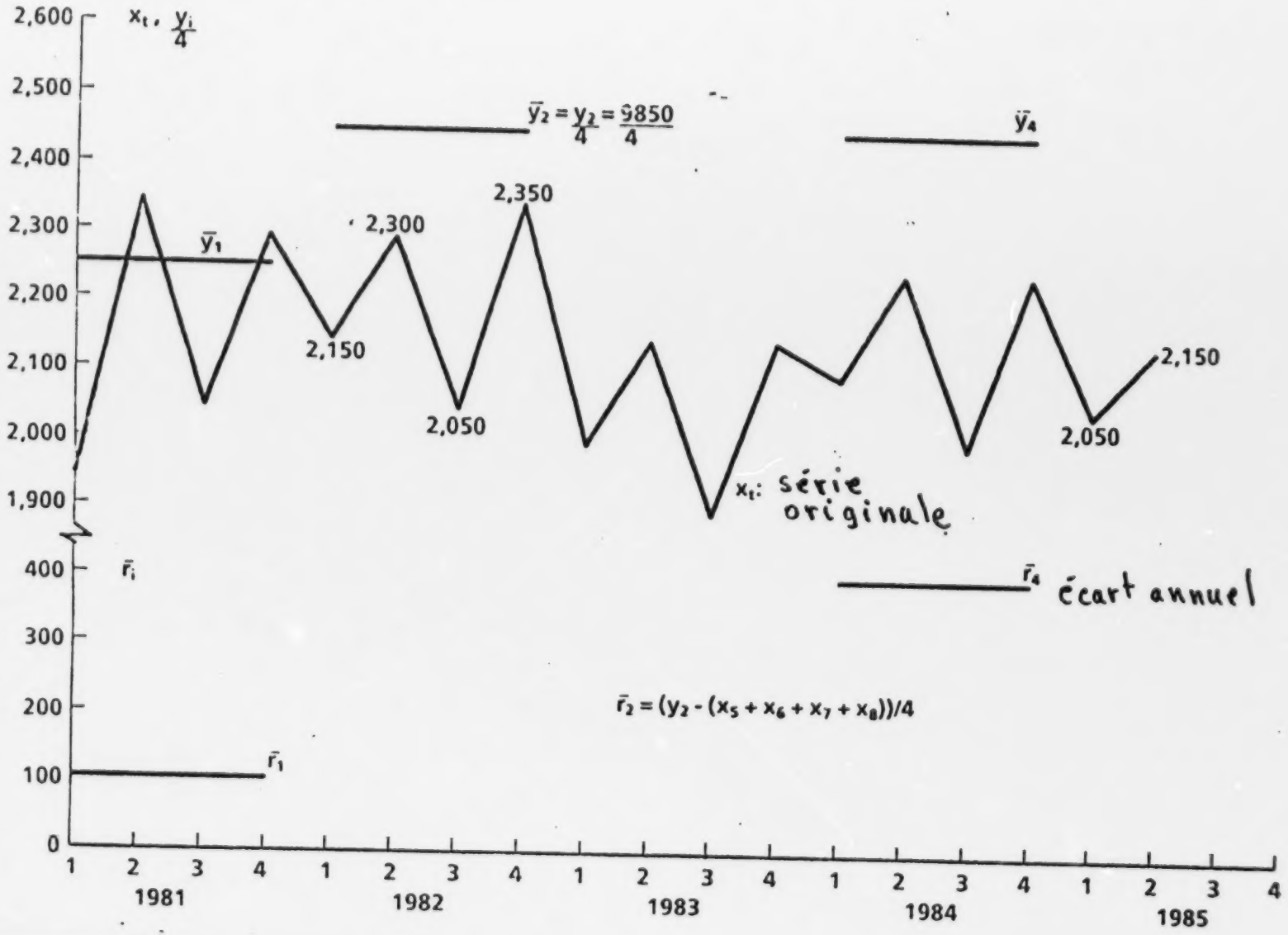
Situation d'étalonnage d'une série de flux



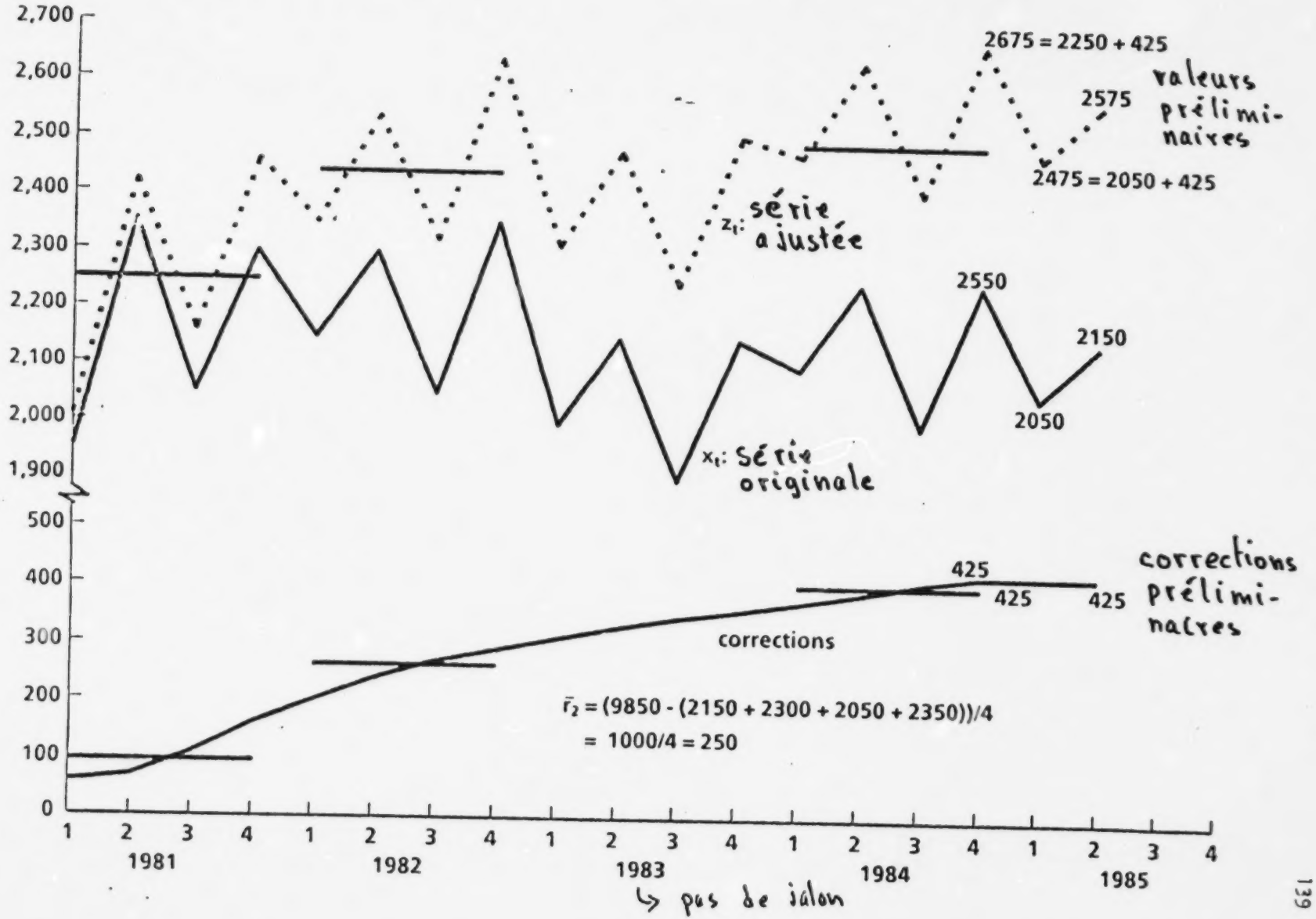
Solution au problème d'étalonnage d'une série de flux



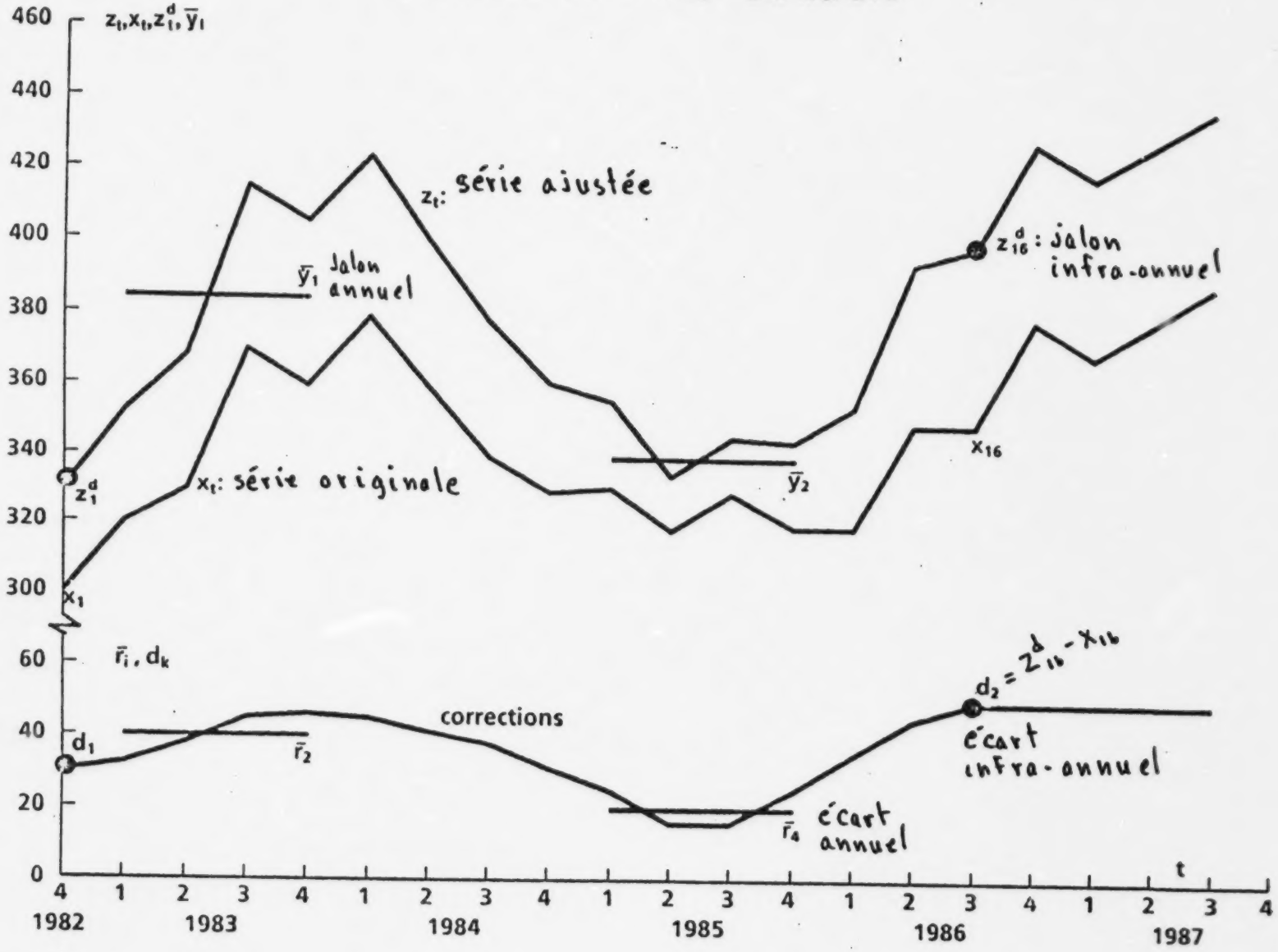
Exercice sur l'étalonnage des séries de flux



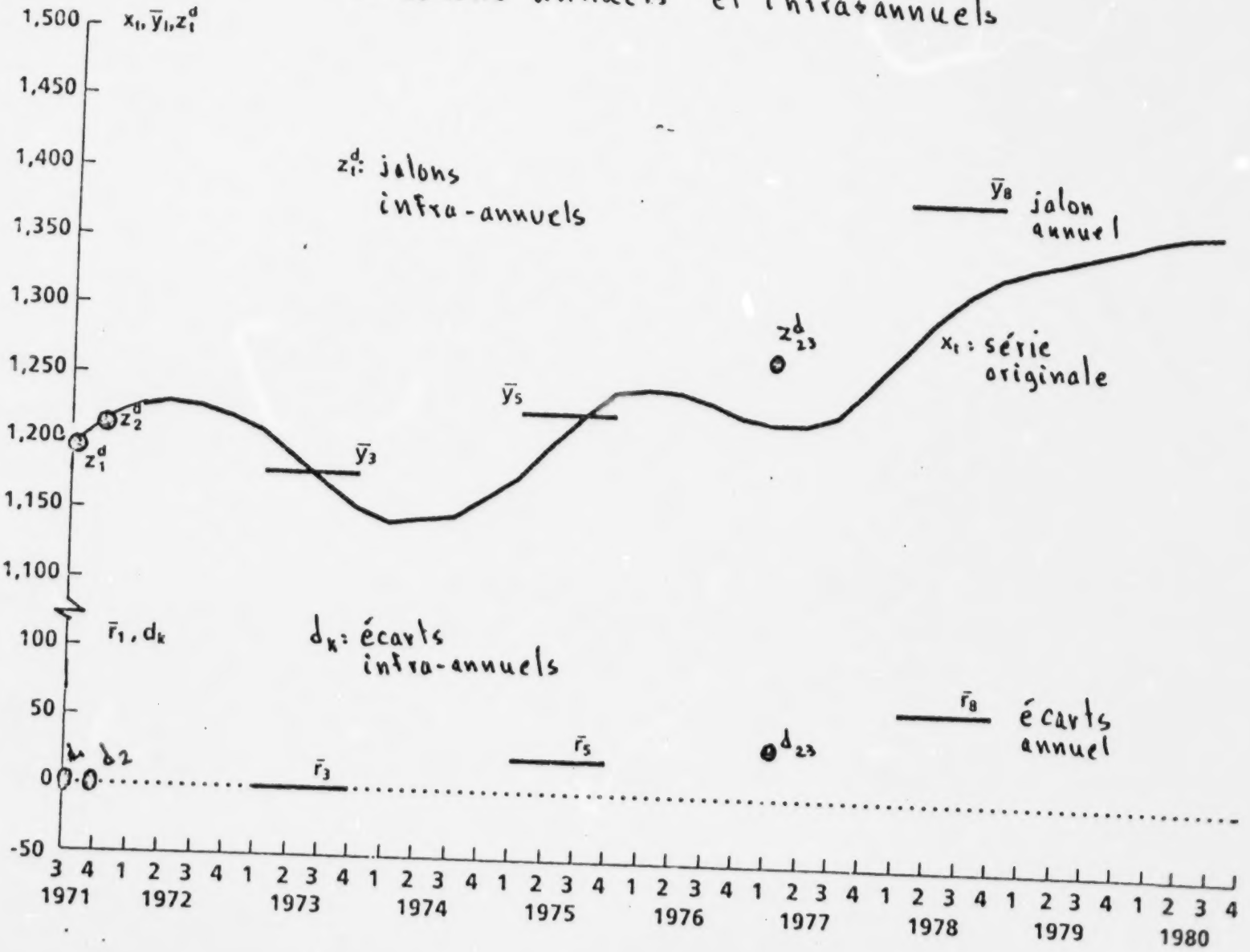
Réponse avec commentaires



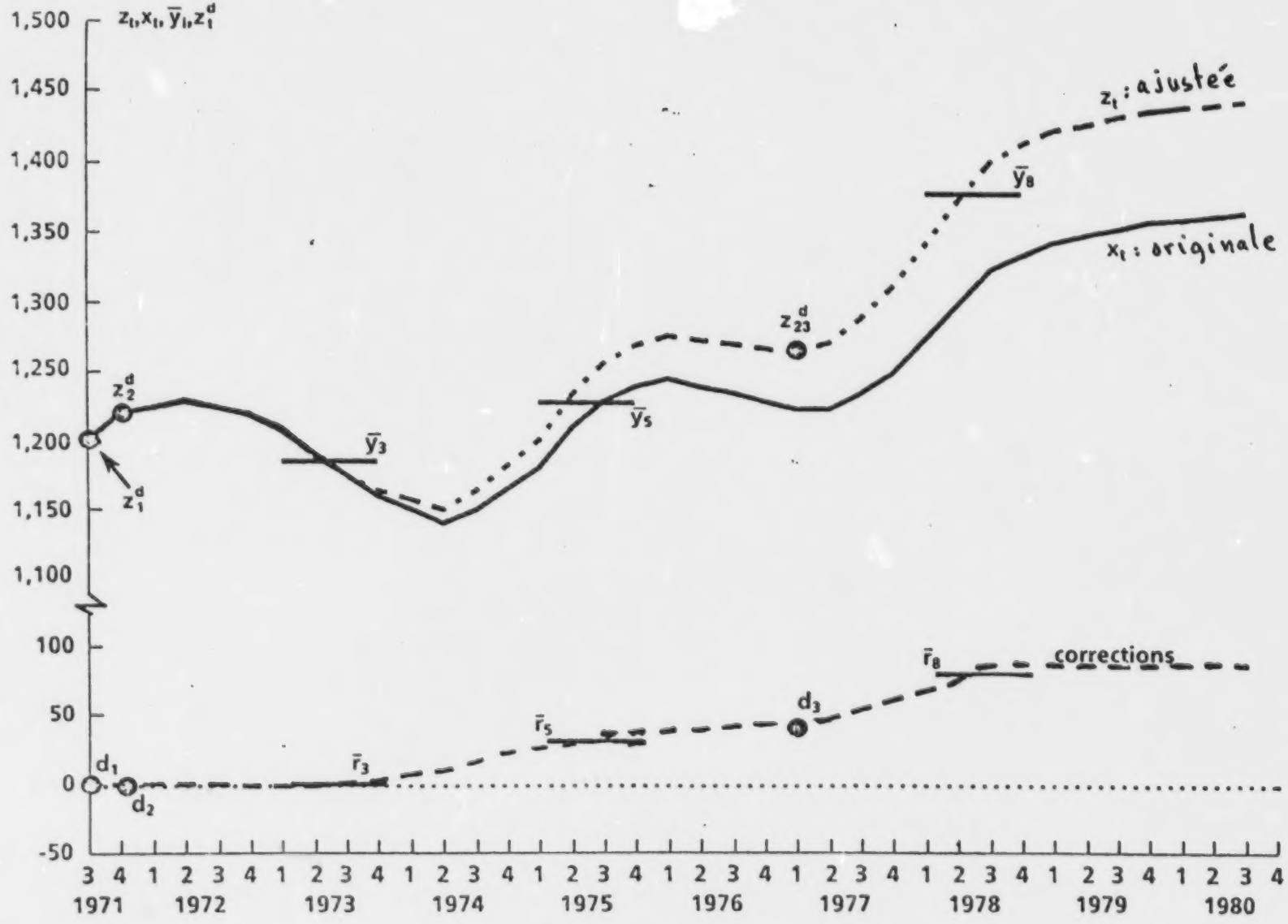
Etalonnage d'une série de flux en présence de jalons annuels et infra-annuels



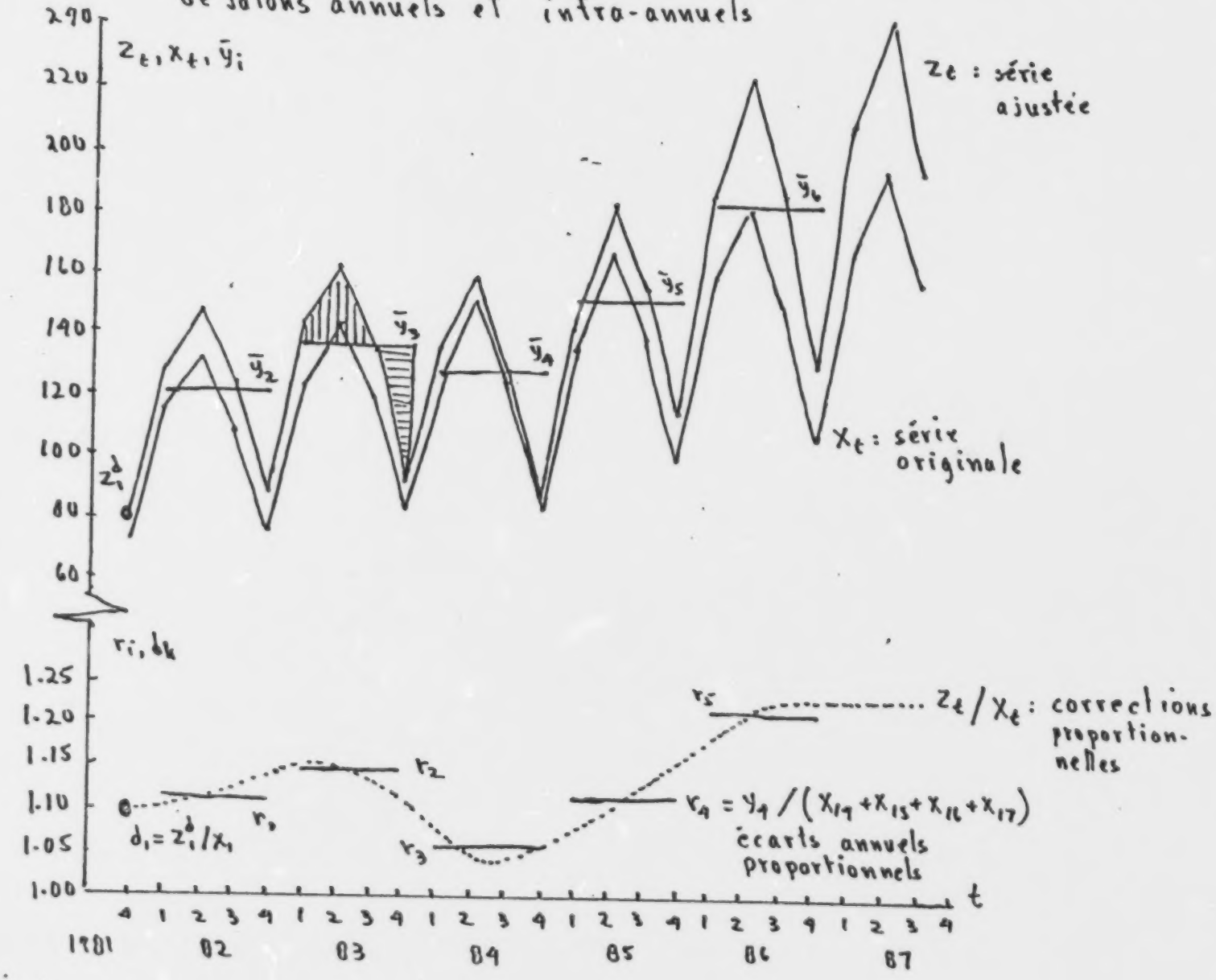
Exercice sur l'étalonnage de séries de flux en présence de jalons annuels et infra-annuels



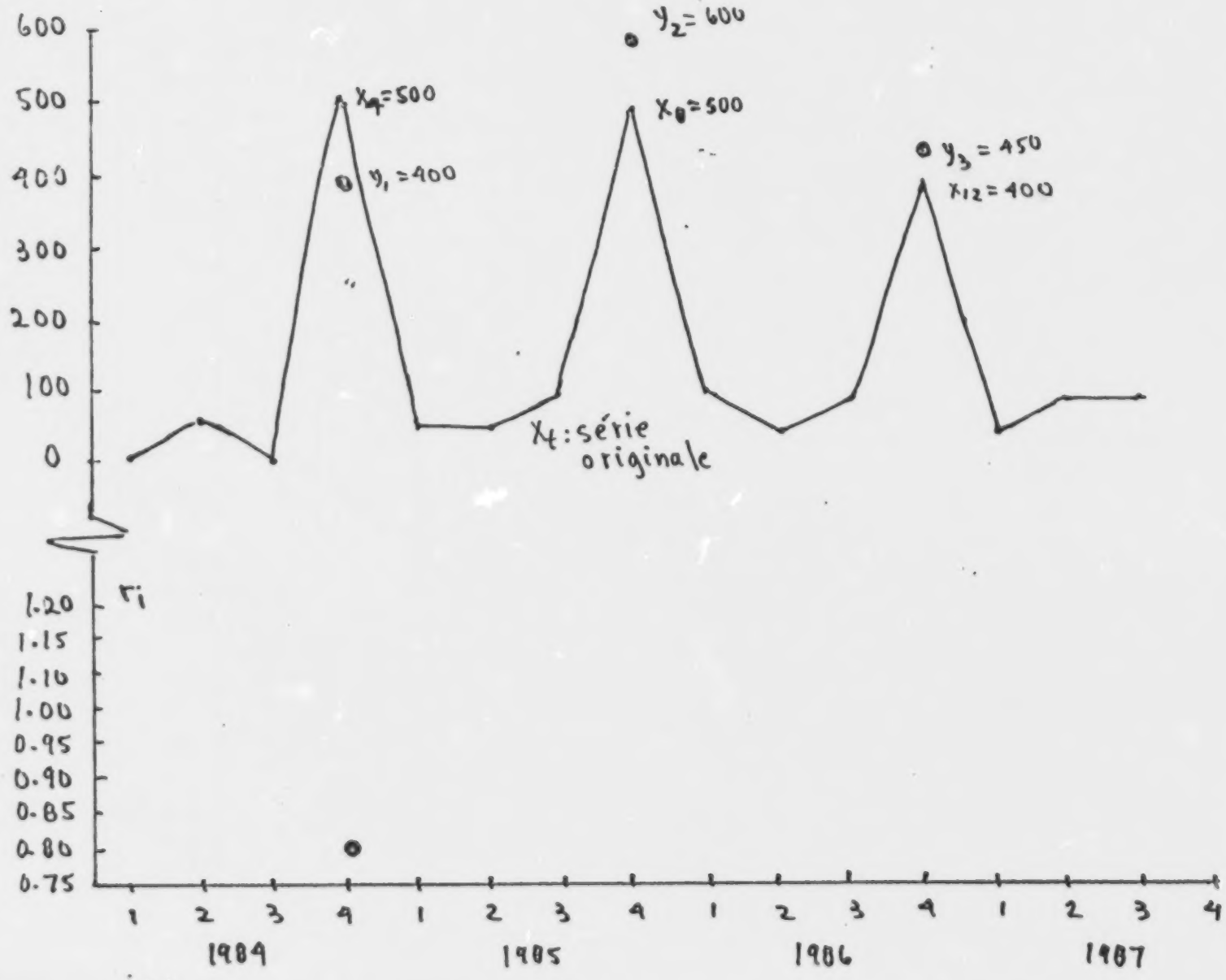
Réponse avec commentaires



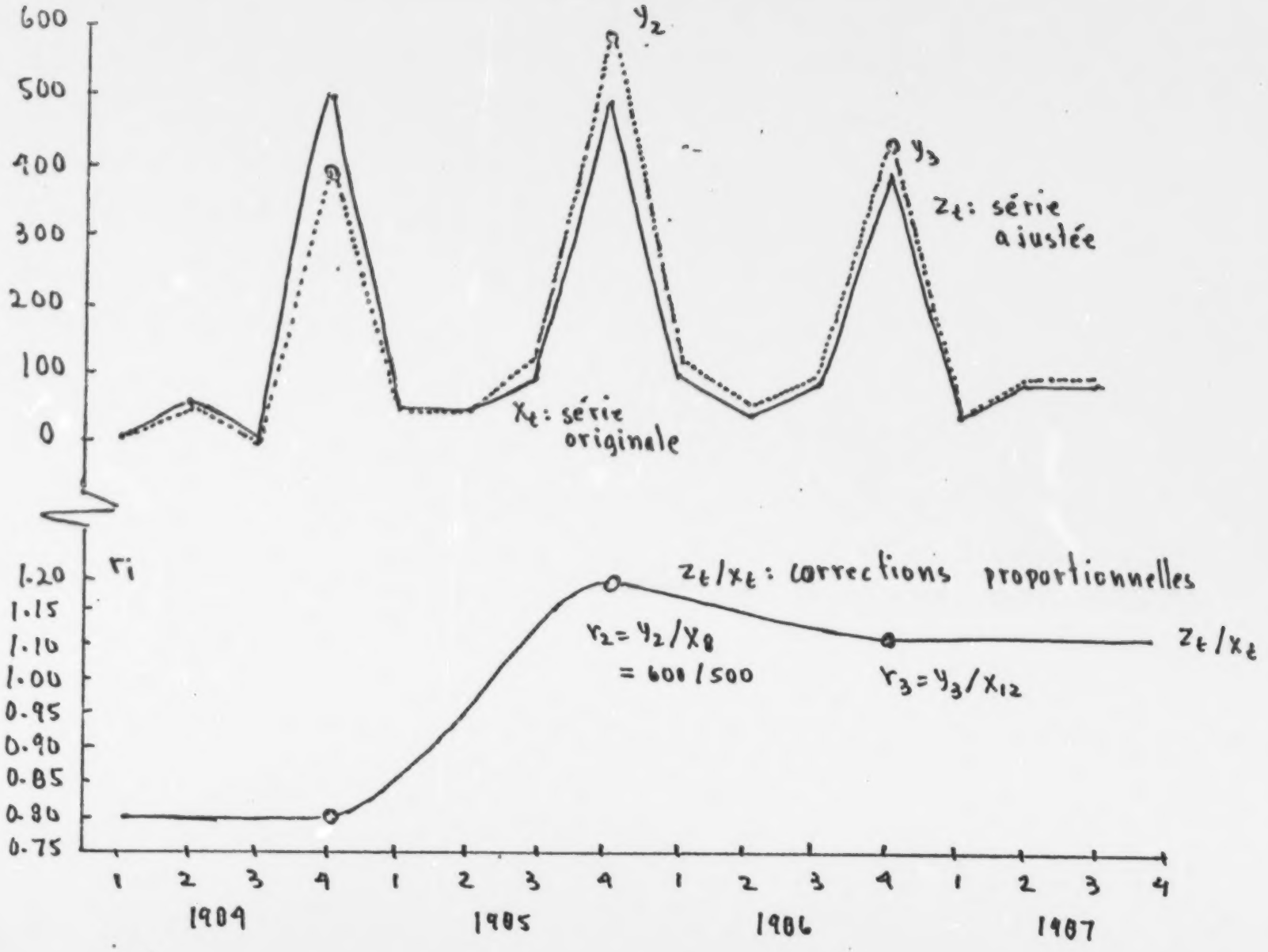
Etalonnage proportionnel d'une série de flux en présence de jalons annuels et infra-annuels



Exercice sur l'étalonnage proportionnel d'une série de stock



Réponse avec commentaires



Mathématique sous-jacente aux situations d'étalonnage examinées

$$\min g(z) = \sum_{t=2}^{T} \left((z_t - z_{t-1}) - (x_t - x_{t-1}) \right)^2 \quad \text{(critère parallélisme)}$$

$$\text{ou bien } \sum_{t=2}^{T} \left(z_t / x_t - z_{t-1} / x_{t-1} \right)^2 \quad \text{(critère paral. proportionnel)}$$

$$+ \sum_{m=1}^{M} w_m \cdot \left(\left(\sum_{j=1}^{J} z_{(i_m - 2)J + J_0 + j} \right) - y_{i_m} \right)^2 \quad \text{(crit. satisfaction des jalons annuels)}$$

$$+ \sum_{k=1}^{K} p_k \left(z_{t_k} - z'_{t_k} \right)^2 \quad \text{(crit. de satisfaction des jalons infra-annuels)}$$

x_t : série originale

z_t : série désirée (inconnue)

T : nbre d'observ. ds serie

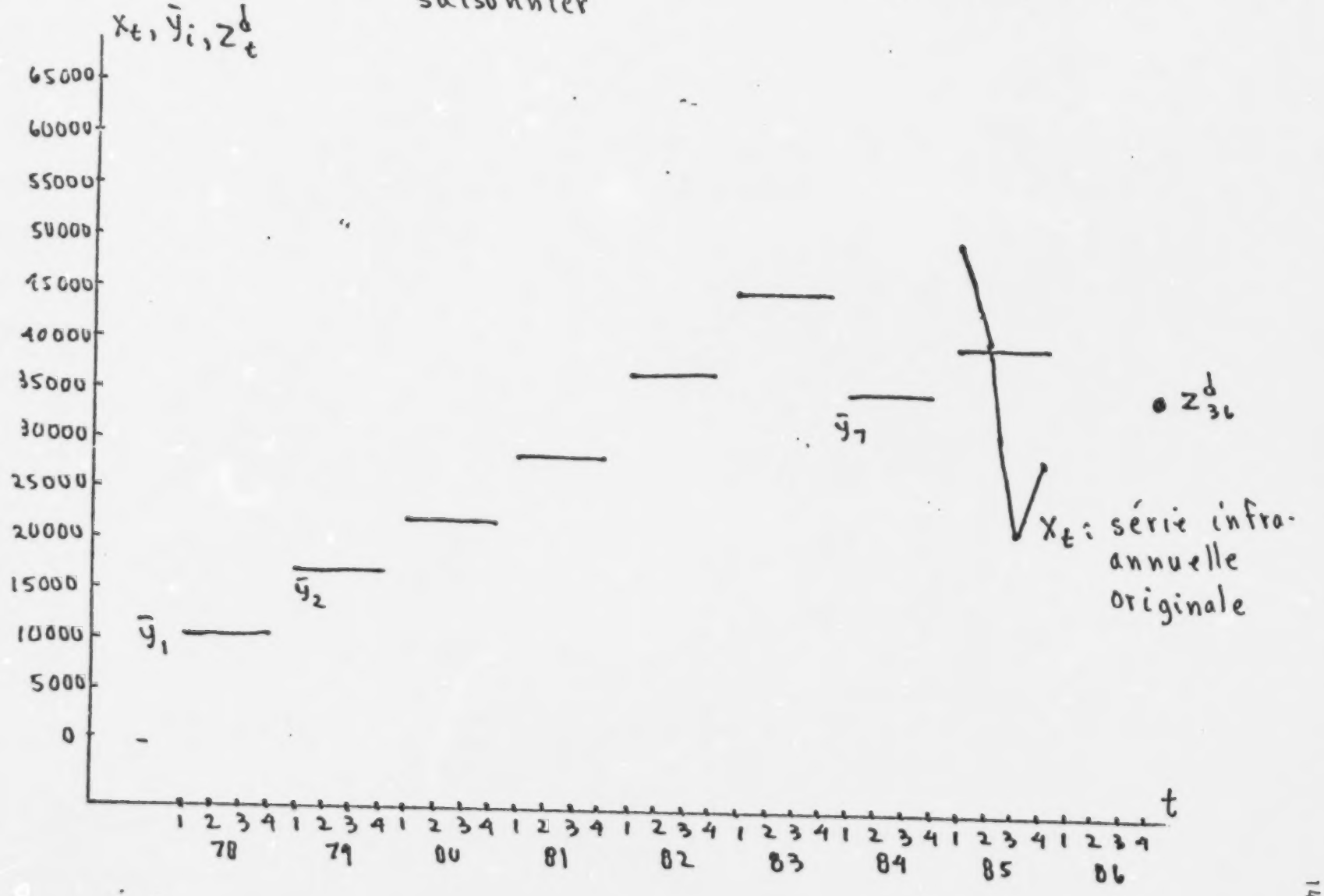
J : nbre de "mois" par "année" ; J_0 : nbre de "mois" ds 1ère année $(0 < J_0 \leq J)$

$i_1, i_2, \ldots, i_M$: "années" de jalons annuels disponibles

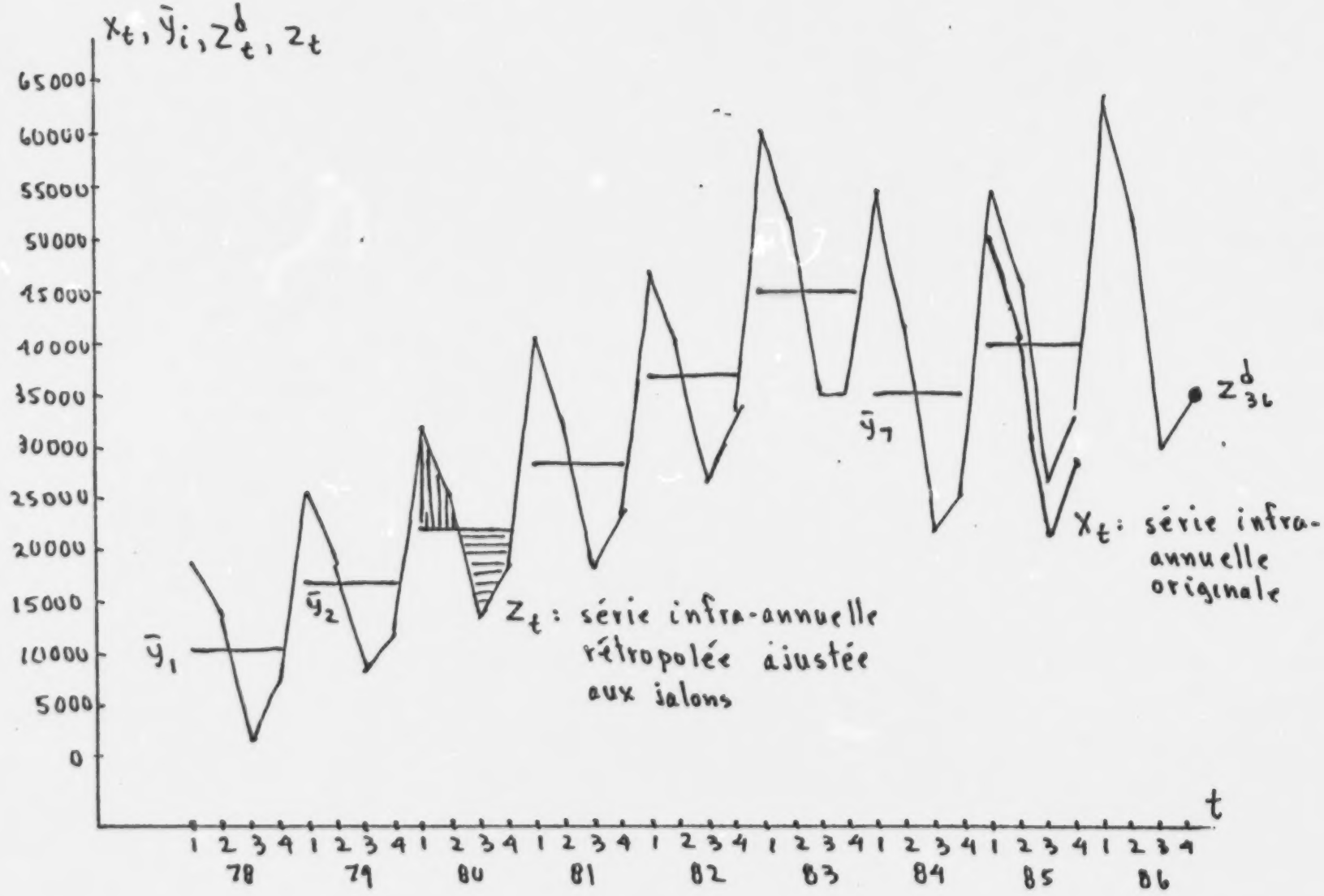
$t_1, t_2, \ldots, t_K$: "mois" de jalons infra-annuels disponibles

w_m, p_k : poids (connus) attribués aux jalons par rapport au critère de parallélisme. (Valeurs élevées ⇒ jalons contraignants)

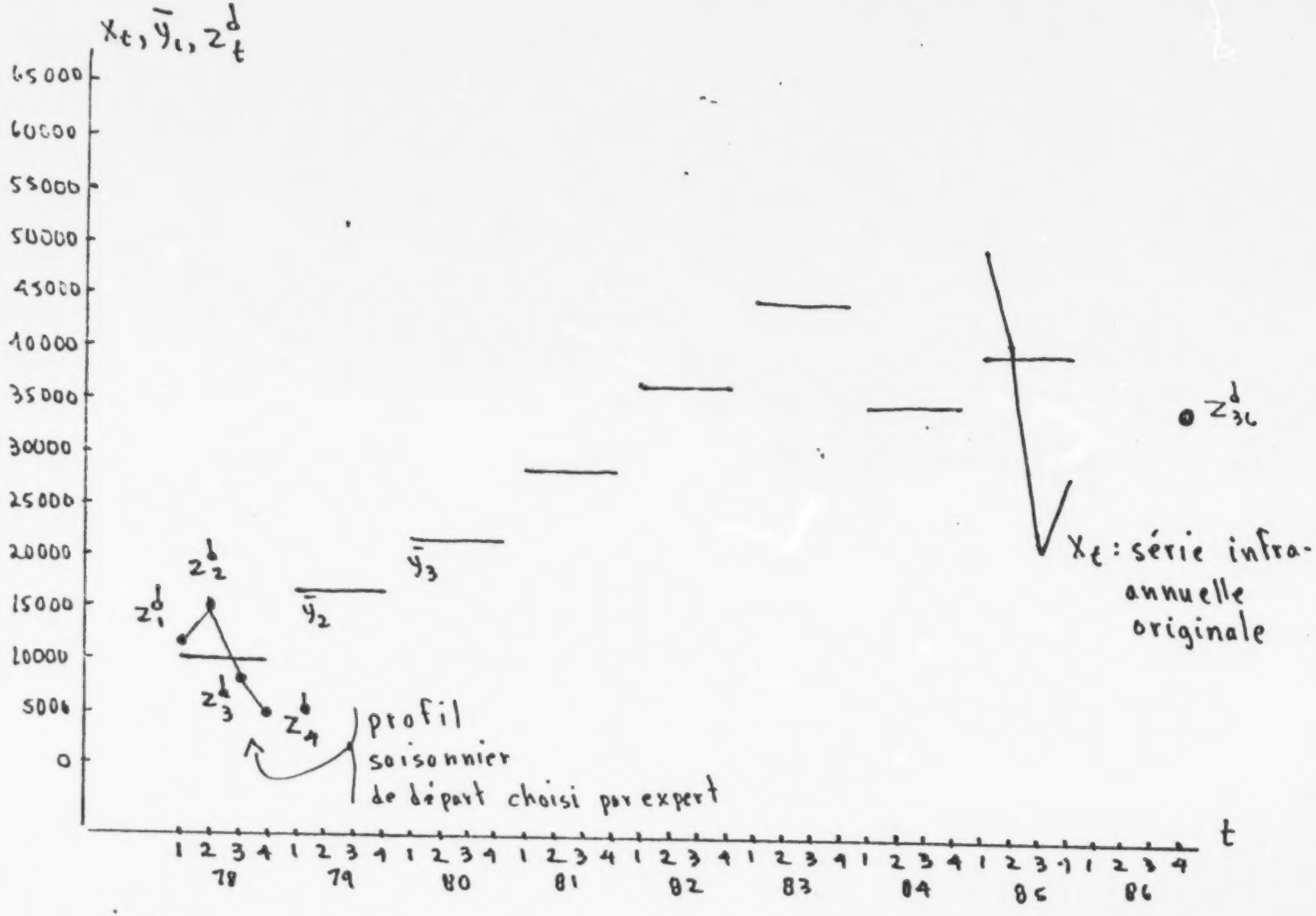
Exercise sur l'interpolation (rétropolation) d'un profil saisonnier



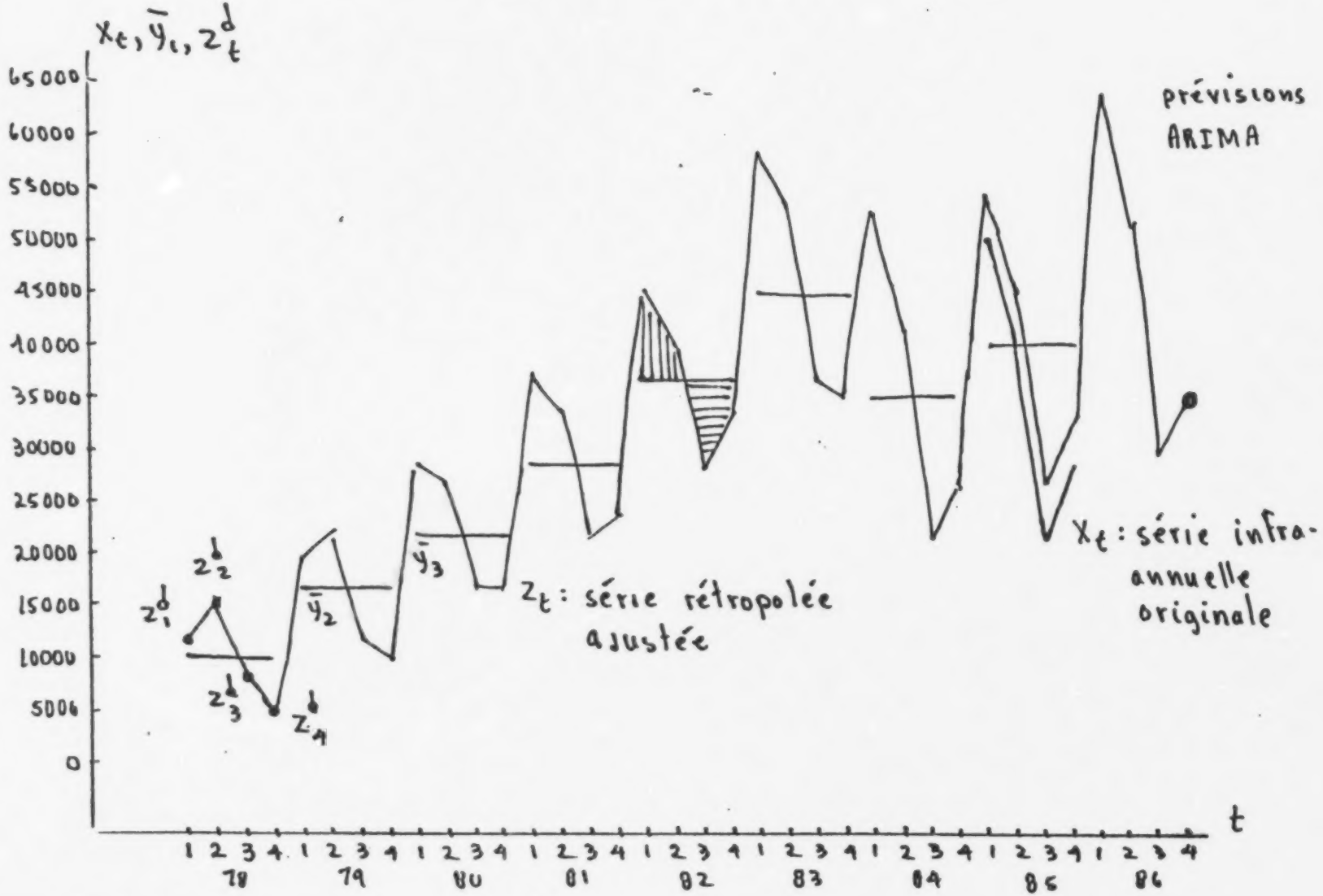
Réponse avec commentaires



Exercise sur l'interpolation entre profils saisonniers



Réponse avec commentaires



Mathématique sous-jacente aux situations d'interpolation examinées

$$\text{Min } g(z) = c \sum_{m=1}^{N} \left((z_{t_m} - z_{t_{m-1}}) - (x_{t_m} - x_{t_{m-1}}) \right)^2 \quad \text{(critère parallélisme)}$$

$$+ (1-c) \sum_{t=J+2}^{T} \left(z_t - \varphi z_{t-1} - \Phi z_{t-J} + \varphi\Phi z_{t-J-1} \right)^2 \quad \text{(critère ARIMA)}$$

$$+ \sum_{m=1}^{M} w_m \left(\left(\sum_{j=1}^{J} z_{(t_m-2)J+J_0+j} \right) - y_{i_m} \right)^2 \quad \text{(crit. satis. jalons annuels)}$$

$$+ \sum_{k=1}^{K} p_k \left(z_{t_k} - z^{d}_{t_k} \right)^2 \quad \text{(crit. satis. jalons infra-annuels)}$$

$t_1, t_2, \ldots t_N$: périodes pour lesquelles le critère de parallélisme est en vigueur

φ, Φ : paramètres autorégressifs connus (pré-déterminés)

Modèle ARIMA saisonnier simples

$$z_t - z_{t-1} = z_{t-J} - z_{t-J-1} + a_t$$

$$z_t - z_{t-1} = \Phi (z_{t-J} - z_{t-J-1}) + a_t$$

$$z_t - \varphi z_{t-1} = \Phi (z_{t-J} - \varphi z_{t-J-1}) + a_t \rightarrow$$

$$z_t - \varphi z_{t-1} - \Phi z_{t-J} + \varphi\Phi z_{t-J-1} = a_t$$

Exercice sur l'établissement de séries à l'aide de taux de croissance et d'une année de base

z^d_t , z_t

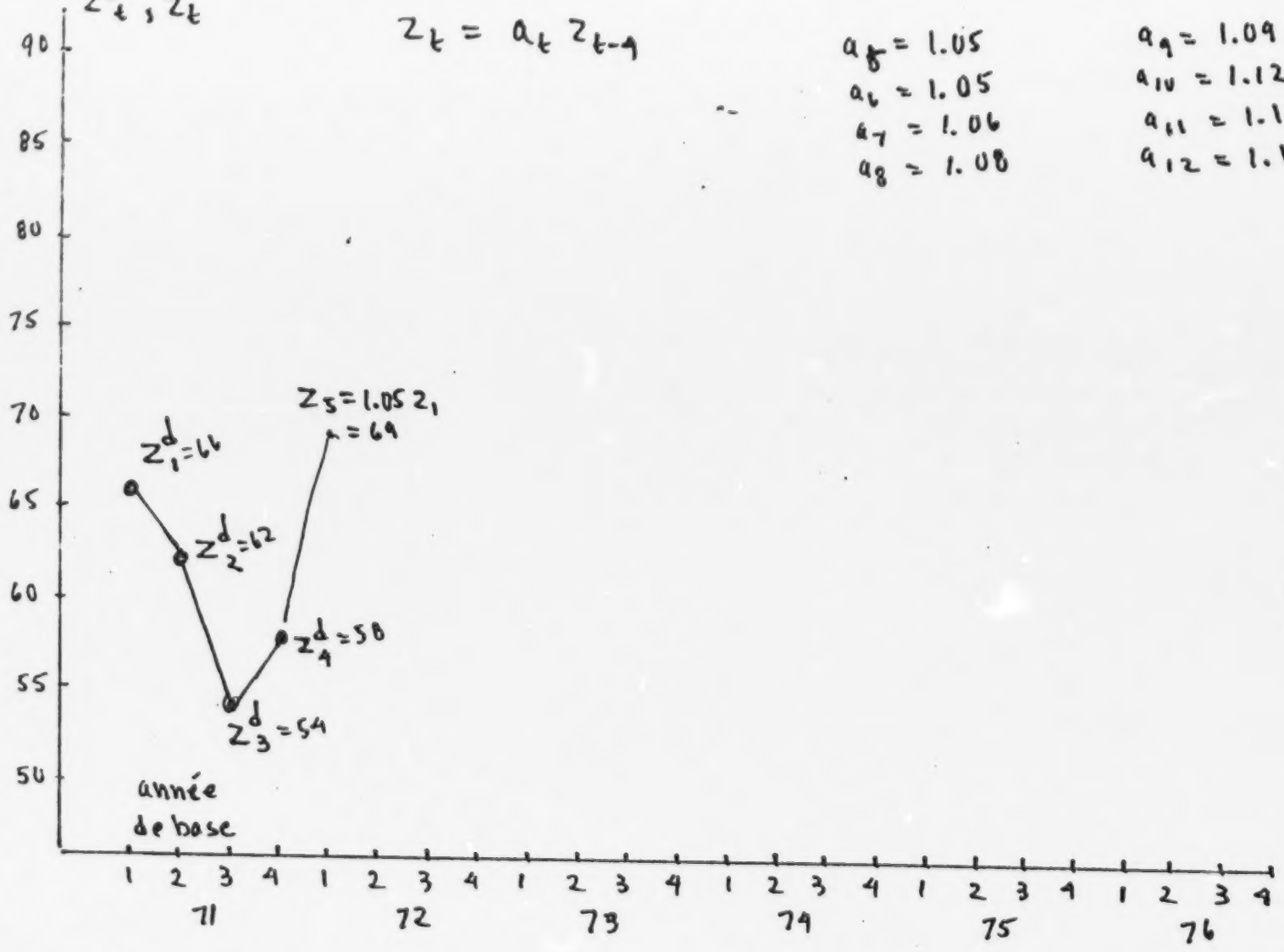
$$z_t = a_t \, z_{t-1}$$

$a_5 = 1.05$ $\quad a_9 = 1.09$

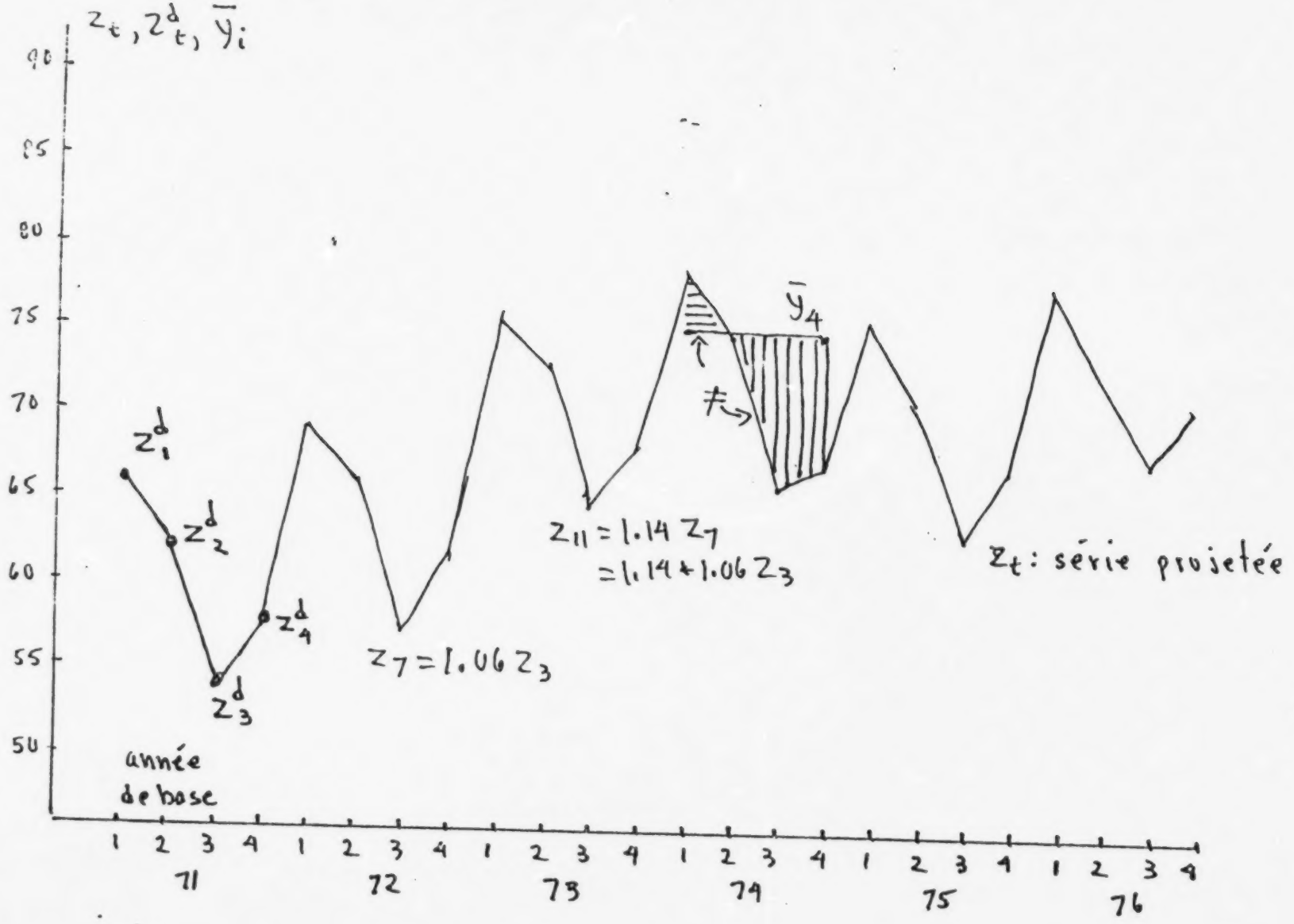
$a_6 = 1.05$ $\quad a_{10} = 1.12$

$a_7 = 1.06$ $\quad a_{11} = 1.14$

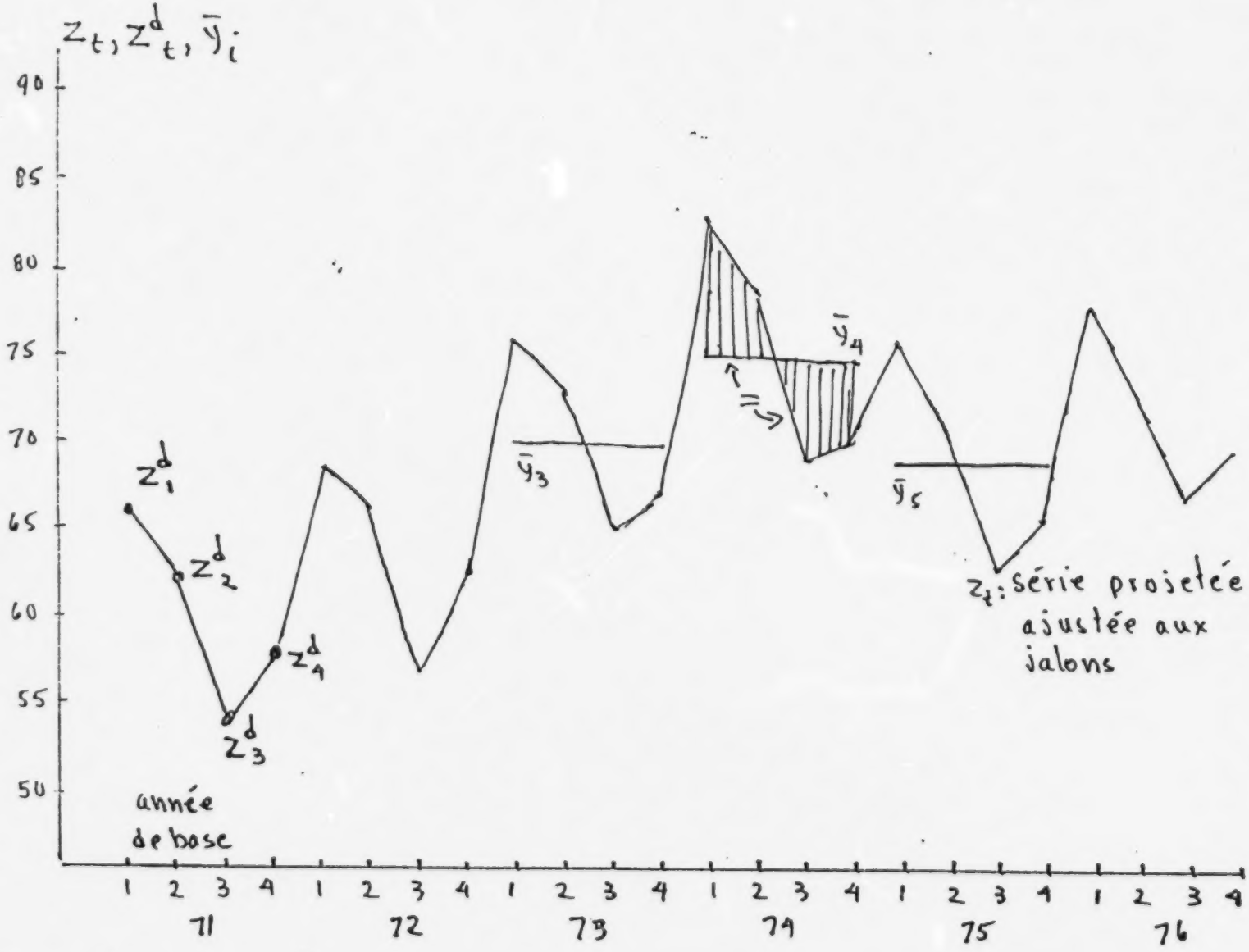
$a_8 = 1.08$ $\quad a_{12} = 1.12$



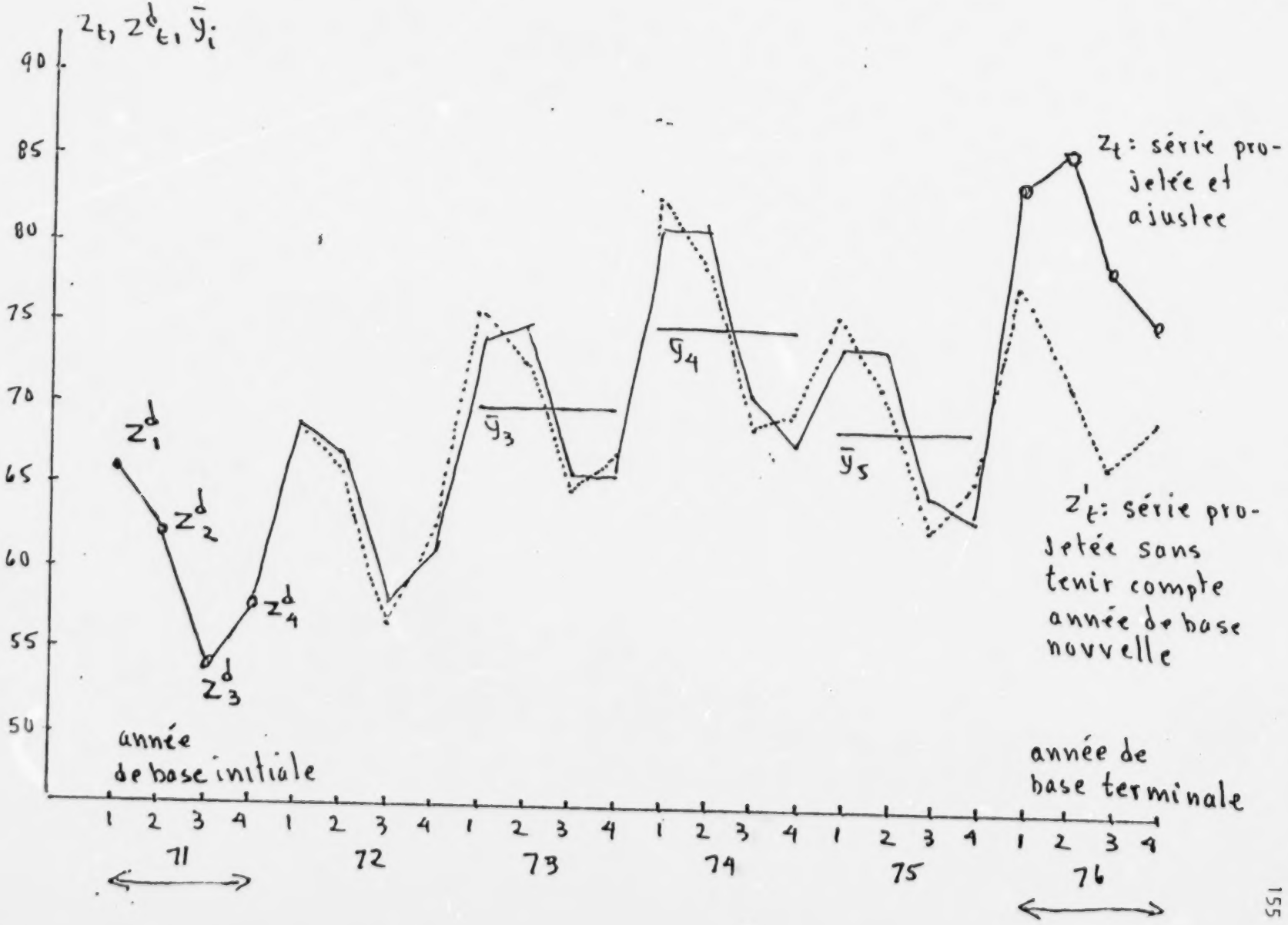
Réponse avec commentaires



Interpolation d'une série à l'aide de taux de croissance et de jalons



Amélioration rétro-active des projections par l'incorporation d'une année de base terminale



La mathématique de l'interpolation à l'aide de taux de croissance et de jalons annuels et infra-annuels

$$\text{Min} \quad g(z) = \sum_{t=J+1}^{T} \left(z_t - a_t z_{t-J} \right)^2$$ (critère de satisfaction des taux de croissance a_t)

$$+ \sum_{m=1}^{M} w_m \left(\left(\sum_{j=1}^{J} z_{(i_m-2)J+J_0+j} \right) - y_{i_m} \right)^2$$ (critère satis. jalons annuels)

$$+ \sum_{k=1}^{K} p_k \left(z_{t_k} - z^d_{t_k} \right)^2$$ (critère satisfaction jalons infra-annuels)

Projection:

$$z_t = a_t z_{t-J} + e_t$$

$$\rightarrow z_t - a_t z_{t-J} = e_t$$

L'estimation robuste des paramètres du modèle AR(1)

CLAUDE LABERGE

1-Introduction

Depuis une dizaine d'années, plusieurs techniques d'estimation robuste ont été mises au point afin d'améliorer la qualité de l'inférence effectuée à partir d'un échantillon contenant des valeurs douteuses. Le manque de robustesse des méthodes d'estimation classiques face à ces valeurs douteuses aura en quelque sorte été le catalyseur ayant permis l'apparition d'un grand nombre de méthodes robustes.

Des méthodes ont d'abord été developpées pour l'estimation d'un paramètre de localisation, par la suite l'estimation de paramètres d'échelle, la regression, l'analyse de variance ont été des domaines pour lesquels des méthodes robustes ont été adaptées afin de tenir compte des valeurs douteuses. Dernièrement des méthodes d'estimation robuste ont été adaptées au domaine des séries chronologiques, ainsi depuis 1979 des auteurs se sont tournés vers ce domaine plus complexe à cause de la dépendance entre les données étudiées.

Le présent exposé se veut un résumé des différentes méthodes qui ont été adaptées au domaine des séries de temps. Afin de simplifier les notations et de permettre une visualisation plus facile des résultats, les développements seront effectués dans le cas d'un processus autorégressif d'ordre 1. Le lecteur intéressé à la généralisation aux modèles ARMA(p,q) peut consulter les articles en référence.

Plusieurs concepts de robustesse ont été présentés dans les dernières années. MARTIN et YOHAI (1985 p.119-126) présentent quatre de ces concepts : la robustesse en terme d'efficacité, la robustesse Minimax, la résistance et la robustesse qualitative. Dans ce travail, la robustesse Minimax ne sera pas discutée. Il sera parfois question de robustesse en terme d'efficacité, on utilise alors les variances asymptotiques comme critère de qualité des estimateurs. Finalement, le concept de robustesse qui nous intéresse le plus ici est celui de résistance présenté ainsi par TUKEY (1976) :

An estimate T_n is called resistant if "small" changes in the data result in only small changes in T_n, where "small changes" in the data means i) large changes in a small fraction of the data, and/or ii) small changes in all the data.

Bien que cette notion de résistance puisse apparaître pratique au statisticien appliqué, MARTIN et YOHAI (1985 p.122) citent le manque de précision de cette définition. Ce concept de résistance a cependant permis de proposer une nouvelle approche à la robutesse qualitative qui apparaît beaucoup plus mathématique. HAMPEL (1971) présente ce concept de robustesse des estimateurs comme une propriété de continuité des fonctionnelles qui les représentent. Etant donné la grande similitude des deux derniers concepts, aucune distinction ne sera faite, entre les deux, dans le présent travail.

2-Modèles de séries avec valeurs aberrantes

En pratique, la construction de modèles de séries chronologiques avec valeurs aberrantes comportent certaines difficultés. En effet, à cause de la présence de non-indépendance des observations, on peut supposer que l'effet d'une valeur aberrante peut être transmis aux observations suivantes à l'aide du caractère autorégressif. Il est aussi concevable que les valeurs aberrantes aient un effet local seulement et n'affectent donc pas le futur du processus. Deux modèles, d'abord présentés par FOX (1972) et repris par DENBY et MARTIN (1979) seront étudiés dans ce travail. Chacun des modèles traitant l'un des deux cas de valeurs aberrantes présentés. Etant donné que ces deux modèles représentent les cas généralement étudiés dans les articles en référence, aucun autre modèle pouvant exister dans la littérature ne sera utilisé.

Modèle avec "Innovation outliers"

Ce premier modèle a la forme :

$$x_t = \phi x_{t-1} + e_t$$

$$y_t = x_t,$$

où $|\phi| < 1$, et où les impulsions e_t sont indépendantes et identiquement distribuées. La fonction de densité $f(\cdot)$ des e_t est généralement supposée symétrique par rapport

à l'origine. Les "innovation outliers" apparaissent alors dans le processus lorsque $f(\cdot)$ est caractérisée par des queues de distributions plus grandes que celles de la loi normale.

On peut, par exemple, supposer que la fonction de densité des impulsions est une normale contaminée :

$$f_\gamma(\cdot;\gamma,\sigma_v^2) = (1-\gamma)\mathcal{N}(\cdot;0,1) + \gamma\mathcal{N}(\cdot;0,\sigma^2)$$

l'utilisation de loi de densité symétrique permet alors de conserver μ comme moyenne générale du processus mesuré.

On peut cependant être interessé à modéliser des valeurs aberrantes de moyennes non-nulles. En supposant que ces valeurs viennent d'une autre population, on a alors que la moyenne du processus mesuré n'est plus équivalente à la moyenne du processus sous-jacent.

Prenons par exemple une loi de densité $f(\cdot)$, normale contaminée :

$$f_\gamma(\cdot;\gamma,\sigma_v^2) = (1-\gamma)\mathcal{N}(\cdot;0,1) + \gamma\mathcal{N}(\cdot;\mu,\sigma^2)$$

où $\mu \neq 0$ et où σ^2 peut être égal à 1. On voit alors que la moyenne du processus mesuré $\{y_t\}_{t=1}^T$ n'est plus égale à la moyenne du processus sous-jacent (sans valeurs aberrantes) qui est toujours égale à zéro.

Voyons quelques exemples permettant de visualiser ces deux cas particuliers. Prenons comme série chronologique mesurée, le cours hebdomadaire de l'action d'une compagnie. Supposons maintenant que la compagnie fasse une annonce importante sur ses résultats financiers. On aura alors généralement pour cette semaine là, une variation beaucoup plus grande du titre mais comme l'annonce peut être bonne ou mauvaise, on peut supposer que la moyenne du processus demeure la même.

Prenons maintenant comme série chronologique mesurée, le débit journalier d'une rivière. Les inondations peuvent alors être vues comme des valeurs aberrantes de moyennes différentes à la moyenne du processus. En effet, comme les inondations

amènent toujours un débit plus grand, elles augmenteront donc la moyenne du processus sous-jacent : les débits journaliers sans valeurs aberrantes.

On peut voir avec ces deux exemples que dans ce premier modèle, l'effet d'une grande impulsion e_t sera propagé à l'aide de la partie autorégressive du modèle (notons de plus que la présentation à l'aide d'un AR(1) ici, peut facilement être généralisée au cas d'un AR(p), MARTIN (1980 p. 228)).

On note que l'effet d'un "innovation outlier" sera propagé même dans le cas d'un MA(q). Cependant, une telle valeur aberrante n'affectera que (q+1) termes dans un MA(q) alors qu'elle affecte toutes les valeurs futures dans un AR(p). Il faut bien sûr noter que l'effet diminuera à mesure que l'on s'éloignera de la valeur en question. Comme l'effet de ces valeurs aberrantes est propagé, on peut alors imaginer que leur origine est quand même interne ou relié au processus sous-jacent.

Modèle avec "Additive outliers"

Le modèle correspondant à ce type de valeur aberrante se présente de la façon suivante:

$$x_t = \phi x_{t-1} + e_t$$

$$y_t = x_t + v_t,$$

où cette fois, la fonction de distribution des e_t est supposée normale de moyenne 0 et de variance σ^2. Les v_t sont également supposés indépendants et identiquement distribués avec distribution :

$$f_\gamma(\cdot;\gamma,\sigma_v^2) = (1-\gamma)\delta(\cdot) + \gamma\mathcal{N}(\cdot;0,\sigma_v^2)$$

où $\delta(\cdot)$ est une loi dégénérée à 0 et où $Pr(v_t = 0) = (1-\gamma)$ et $Pr(v_t \neq 0) = \gamma$. On utilise généralement γ plus petit que 0.2 puisque l'on suppose que les valeurs aberrantes sont peu nombreuses.

On peut de nouveau être interessé à modéliser des valeurs aberrantes de moyennes non-nulles. On peut alors utiliser comme fonction de densité des v_t :

$$f_\gamma(\cdot;\gamma,\sigma_v^2) = (1-\gamma)\delta(\cdot) + \gamma\mathcal{N}(\cdot;\mu,\sigma_v^2)$$

On voit ici que la présence d'un "additive outlier" n'affectera pas les valeurs futures. On peut aussi penser que ces valeurs aberrantes sont plutôt externe au processus ce qui amènerait la raison pour laquelle leur effet n'est pas propagé par le processus.

Voyons quelques exemples : prenons d'abord comme série chronologique mesurée, toute mesure effectuée en laboratoire. Les erreurs de mesures ou de transcription peuvent alors être vues comme valeurs aberrantes de plus ou moins grande amplitude. Comme ces erreurs peuvent augmenter ou diminuer la moyenne, on peut donc supposer qu'elles ont la même moyenne que le processus tout en introduisant généralement une augmentation de la variance.

Prenons maintenant comme série chronologique mesurée, un paramètre de qualité de l'eau d'une rivière. Les déchargements de produits polluants en amont immédiat du site d'échantillonnage s'ils ne sont pas permanents, peuvent alors être vus comme des valeurs aberrantes. Comme ces déchargements amènent toujours une détérioration de la qualité de l'eau, on doit alors supposer que la moyenne des valeurs affectées par ces rejets est différente à la moyenne du processus sous-jacent.

3-Estimateurs et leurs caractéristiques

A) Estimations classiques.

Dans le cas d'un AR(1) avec $\mu \neq 0$:

$$(y_t - \mu) = \phi(y_{t-1} - \mu) + e_t, \qquad t = 1, \ldots, T,$$

trois paramètres doivent être estimés, il s'agit de μ : la moyenne du processus, σ_e^2: la variance des impulsions (qui sera notée simplement σ^2 dans le reste de ce chapitre), et ϕ : le coefficient autorégressif d'ordre 1. Les méthodes d'estimations classiques ont été étudiées par plusieurs auteurs. Par exemple, BUSTOS et AHSANULLAH (1984) présentent les résultats de l'estimation du maximum de vraisemblance pour les estimateurs de ϕ et de σ^2. Cette méthode utilise une fonction de vraisemblance construite à partir de la fonction de distribution normale bivariée de y_t et y_{t-1}:

$$\ln L = \frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 + \frac{1}{2}\ln(1-\phi^2) - (2\sigma^2)^{-1}\left[(1-\phi^2)(y_1-\mu) + \sum_{t=2}^{T}[(y_t-\mu)-\phi(y_{t-1}-\mu)]^2\right]$$

L'expression des estimateurs est alors assez compliquée et la solution, des équations provenant des dérivées partielles de cette fonction de vraisemblance, n'est pas unique. Ces estimateurs ne seront pas utilisés dans les pages qui suivent et ne seront donc pas présentés ici. Le lecteur désirant en savoir plus, peut consulter BUSTOS et ASHANULLAH (1984) ou encore FULLER (1976 p.328).

Une méthode d'estimation beaucoup plus utilisée pour ce problème est la construction d'une fonction de vraisemblance conditionnelle. On suppose alors y_1 connu et on minimise la fonction de vraisemblance de $y_2 \ldots y_T$ étant donné y_1:

$$\ln L = -\frac{T-1}{2}\ln 2\pi - \frac{T-1}{2}\ln\sigma^2 - (2\sigma^2)^{-1}\sum_{t=2}^{T}[(y_t-\mu)-\phi(y_{t-1}-\mu)]^2 \qquad 3.A.1$$

Les estimateurs alors obtenus sont les suivants:

$$\hat{\phi} = \frac{\sum_{t=2}^{T}(y_t-\hat{\mu})(y_{t-1}-\hat{\mu})}{\sum_{t=2}^{T}(y_{t-1}-\hat{\mu})^2}$$

$$\hat{\mu} = \frac{T}{T-1)}\bar{y} + \frac{\hat{\phi}y_T - y_1}{(1-\hat{\phi})(T-1)}$$

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=2}^{T}[(y_t-\hat{\mu})-\hat{\phi}(y_{t-1}-\hat{\mu})]^2$$

On peut remarquer que ces estimateurs sont très semblables aux estimateurs des moindres carrés obtenus en minimisant:

$$S^2 = \sum_{t=2}^{T}[(y_t-\mu)-\phi(y_{t-1}-\mu)]^2 \qquad (3.A.2)$$

par rapport à ϕ et μ . En fait ils sont identiques pour l'estimation de μ et de ϕ puisque la partie de la fonction de vraisemblance conditionnelle qui dépend de ϕ et μ est égale à une constante multiplicative près à S^2. Pour σ^2, l'estimateur de la variance des impulsions, la différence est de l'ordre de $o(T^{-\frac{1}{2}})$, ((T-2) remplaçant (T-1)

au dénominateur). Etant donné cette équivalence presque complète, les termes estimateurs des moindres carrés (EMC) et estimateurs du maximum de vraisemblance conditionnel (EMVC) seront utilisés sans distinction.

La distribution asymptotique de ces estimateurs, sous l'hypothèse de normalité des impulsions, est présentée dans BOX et JENKINS (1976 appendice A7.5) ainsi que dans MARTIN et YOHAI (1985 p.127). L'indépendance asymptotique des estimateurs permet la présentation suivante sous forme de trois lois univariées :

$$T^{\frac{1}{2}}(\widehat{\mu}_{MC} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{\sigma^2}{(1-\phi)^2}) \tag{3.A.3}$$

$$T^{\frac{1}{2}}(\widehat{\phi}_{MC} - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1-\phi^2)) \tag{3.A.4}$$

$$T^{\frac{1}{2}}(\widehat{\sigma}^2_{MC} - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\sigma^4) \tag{3.A.5}$$

Voici les caractéristiques particulières de ces estimateurs dans l'estimation des paramètres d'un processus AR(1). Ces caractéristiques peuvent dans la plupart des cas être facilement généralisées aux processus plus complexes.

a) Les EMVC pour μ et ϕ ne sont pas robustes en terme d'efficacité en présence de "innovation outliers" (MARTIN et YOHAI 1985 p.129) et de "additive outliers" (DENBY et MARTIN 1979 p.141).

b) Le EMVC pour ϕ est convergent en présence de "innovation outliers", pour toute loi $f(\cdot)$ ayant une variance finie (DENBY et MARTIN 1979 p.140), mais ne l'est pas en présence de "additive outliers" (DENBY et MARTIN 1979 p.141).

c) Les EMVC ne sont pas résistants en présence de "additive outliers" et de "innovation outliers" (BUSTOS et YOHAI 1985 p.155).

d) MARTIN et YOHAI (1985 p.128), cite une autre propriété intéressante de l'EMVC pour ϕ ; soit le fait que cet estimateur a une variance asymptotique qui ne dépend pas de la variance des impulsions. Ces auteurs parlent alors de "distribution free property" de l'estimateur de ϕ sous la restriction $Var(F) < \infty$.

Afin d'obtenir une estimation plus robuste pour les paramètres des processus ARMA, certains auteurs se sont alors tourné vers les méthodes proposées par HUBER.

B) M-Estimateurs

MARTIN et DENBY (1979) ont étudié l'estimation des paramètres d'un AR(1) à l'aide des M-Estimateurs de HUBER. Notons cependant que dans plusieurs ouvrages traitant de l'estimation robuste en série de temps, on traite surtout l'estimation du paramètre ϕ pour passer souvent très rapidement sur l'estimation de la moyenne et de la variance. Ainsi, il n'y a pas toujours concensus quant à la façon de traiter ces dernières estimations. Pour ce qui est de l'estimation de la moyenne, on propose souvent de l'estimer à l'aide des méthodes robustes de HUBER ou de TUKEY pour ensuite travailler avec des observations centrées afin de ne plus tenir compte du paramètre de localisation. Ainsi, on propose rarement une méthode d'estimation simultanée des deux paramètres μ et ϕ . Quant au paramètre σ^2, on s'attarde peu à la question de son estimation dans les ouvrages d'estimation robuste en série de temps.

On obtient les M-Estimateurs en minimisant la fonction suivante par rapport aux paramètres en question:

$$\sum_{t=2}^{T} \rho\left(\frac{(y_t - \mu) - \phi(y_{t-1} - \mu)}{\sigma}\right) \tag{3.B.1}$$

où $\rho(\cdot)$ est une fonction arbitraire. Ce problème est équivalent à trouver ϕ_M et μ_M solutions de :

$$\sum_{t=2}^{T} (y_{t-1} - \mu)\psi(b_t) = 0 \tag{3.B.2}$$

$$\sum_{t=2}^{T} \psi(b_t) = 0 \tag{3.B.3}$$

où $b_t = \frac{(y_t-\mu)-\phi(y_{t-1}-\mu)}{\sigma}$: représente l'impulsion centrée et réduite, et où les fonctions ψ $(\psi = \rho')$ généralement utilisées sont :

Fonction de HUBER

$$\psi_{H,k}(v) = \begin{cases} v, & \text{si } |v| < k \\ k\,sign(v), & \text{si } |v| > k \end{cases} \tag{3.B.4}$$

Fonction biweight

$$\psi_{M,k}(v) = \begin{cases} v(1-(\frac{v}{k})^2)^2, & \text{si } |v| < k \\ 0, & \text{si } |v| > k \end{cases} \tag{3.B.5}$$

La fonction de HUBER est une fonction monotone croissante alors que la fonction biweight est une fonction appelée redescendante qui amène une non unicité des solutions de (3.B.2) et (3.B.3). Il sera alors important de choisir une "bonne" valeur initiale lorsque l'on utilisera les algorithmes numériques pour résoudre ce système d'équation.

En ce qui concerne l'estimation de σ, on utilise encore une fois une approche proposée par HUBER (1981) qui consiste à résoudre simultanément l'équation :

$$\sum_{t=2}^{T} \delta(b_t) = 0$$

où $\delta(\cdot)$ est une fonction paire et monotone non-décroissante. HUBER (1981) propose de choisir la fonction $\delta(\cdot)$ dans la famille :

$$\delta(v) = \psi^2(v) - b \tag{3.B.6}$$

avec $b = E_Z(\psi^2(v))$ où Z est une loi $\mathcal{N}(0,1)$. Ce choix de b détermine un estimateur de σ qui est convergent lorsque les impulsions sont normales centrées-réduites.

Les distributions asymptotiques des M-Estimateurs sont traitées rapidement dans MARTIN et YOHAI (1985 p.132). Voyons, à titre de comparaison avec la distribution asymptotique des estimateurs des moindres carrés, les résultats obtenus pour les M-Estimateurs sous l'hypothèse que les impulsions suivent une loi $\mathcal{F}$. Encore une fois, l'indépendance asymptotique des estimateurs permet une présentation sous forme de trois lois univariées :

$$T^{\frac{1}{2}}(\widehat{\mu}_M - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{\sigma^2 E_F(\psi^2(b_t))}{(1-\phi)^2 E_F(\psi'(b_t))^2}) \tag{3.B.7}$$

$$T^{\frac{1}{2}}(\widehat{\phi}_M - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (1-\phi^2)\frac{E_F(\psi^2(b_t))}{E_F(\psi'(b_t))^2}) \tag{3.B.8}$$

$$T^{\frac{1}{2}}(\widehat{\sigma}_M - \sigma) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{E_F(\psi^4(b_t)) - E_F(\psi^2(b_t))^2}{E_F(\psi'^2(b_t))}) \tag{3.B.9}$$

Les études effectuées en grande partie par MARTIN montrent les caractéristiques suivantes pour l'utilisation des M-Estimateurs en série de temps :

a) Le M-Estimateur pour ϕ est robuste au point de vue de l'efficacité en présence de "innovation outliers" avec un bon choix de fonction ψ (DENBY et MARTIN 1979 p.141).

b) Le M-Estimateur pour ϕ est fortement biaisé en présence de "additive outliers", le biais asymptotique obtenu peut être aussi catastrophique que dans le cas de l'estimation à l'aide des moindres carrés (DENBY et MARTIN 1979 p.141).

c) Comme un petit nombre de valeurs peut encore avoir une grande influence sur l'estimation de ϕ par le terme $(y_{t-1} - \mu)$ de l'équation (3.B.2), le M-Estimateur pour ϕ n'est donc pas résistant (MARTIN et YOHAI 1985 p.133).

Il apparaît donc que l'utilisation des M-Estimateurs ne règle pas le problème de non-robustesse de l'estimation des paramètres d'un AR(1).

C) GM-Estimateurs ("Generalized" M-Estimateurs)

Ces estimateurs ont été introduits pour deux raisons: premièrement, réduire le biais en présence de "additive outliers" et deuxièment, rendre l'estimation des paramètres robuste qualitativement. Dans le cas d'un AR(1), le problème a été traité par DENBY et MARTIN (1979) et repris avec certaines modifications dans MARTIN et YOHAI (1985). On introduira ici les GM-Estimateurs d'une façon un peu différente.

On peut d'abord noter que la non-robustesse du M-Estimateur de ϕ vient du fait que même si la fonction $\psi(\cdot)$ est bornée (équation 3.B.2), une petite fraction des

données peut encore avoir une grande influence sur l'estimation car la composante $(y_{t-1}-\mu)$ n'est pas bornée.

On remplacera donc ici l'équation (3.B.2) par :

$$\sum_{t=2}^{T} \eta\left(b_t, \frac{(y_{t-1}-\mu)}{\sigma_y}\right) = 0 \tag{3.C.1}$$

où σ_y représente l'écart-type des observations y_t et où $\eta(\cdot,\cdot)$ est une fonction de poids de $\Re^2$ dans $\Re$ ayant pour but de réduire l'effet d'un résidu de prédiction trop grand en raison d'un "additive outlier" apparaissant dans la variable de prédiction y_{t-1}. Spécifiquement on utilise à cette fin l'une des deux fonctions suivantes :

Fonction de MALLOWS :

$$\eta(v,w) = \psi_1(v)\psi_2(w) \tag{3.C.2}$$

où $\psi_1(t)$ et $\psi_2(t)$ sont des fonctions bornées comme pour les M-Estimateurs.

Fonction de HAMPEL-KRASKER et WELSCH :

$$\eta(v,w) = \psi(vw). \tag{3.C.3}$$

Alors $\hat{\phi}_{GM}$ est la solution de:

$$\sum_{t=2}^{T} \psi_1(b_t)\psi_2\left(\frac{y_{t-1}-\mu}{\sigma_y}\right) = 0 \tag{3.C.4}$$

en utilisant la fonction de MALLOWS et de :

$$\sum_{t=2}^{T} \left(b_t \frac{(y_{t-1}-\mu)}{\sigma_y}\right) = 0 \tag{3.C.5}$$

en utilisant la fonction de HAMPEL-KRASKER et WELSCH.

Pour ce qui est de l'estimation de μ et de σ, les équations (3.B.3) et (3.B.6) sont de nouveau utilisées puisqu'elles définissaient déjà des estimateurs robustes.

En utilisant la fonction de MALLOWS et en supposant σ_y connu, on peut montrer que la distribution asymptotique de l'estimateur $\hat{\phi}_{GM}$ est, sous l'hypothèse que les impulsions suivent une loi $\mathcal{F}$:

$$T^{\frac{1}{2}}(\widehat{\phi}_{GM} - \phi) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, (1-\phi^2)\frac{E_F(\psi^2(b_t))E_F(\psi^2(a_{t-1}))}{E_F(\psi'(b_t))^2 E_F(a_{t-1}\psi(a_{t-1}))^2}\right) \qquad (3.C.6)$$

où $a_{t-1} = \frac{(y_{t-1}-\mu)}{\sigma_y}$.

Voyons maintenant les caractéristiques de ces estimateurs :

a) Comme la variance asymptotique du GM-Estimateur pour ϕ est plus grande que celle du M-estimateur correspondant, il en resulte donc une perte d'efficacité pour ces estimateurs en présence de "innovation outliers" (DENBY et MARTIN 1979 p.142).

b) Le biais asymptotique des GM-Estimateurs est plus faible que le biais des M-Estimateurs en présence de "additive outliers" (DENBY et MARTIN 1979 p.142).

c) Les GM-Estimateurs sont convergents en présence de "innovation outliers" sous des conditions raisonnables de régularité (DENBY et MARTIN 1986 p.141).

d) Dans le cas où le processus contient seulement une partie autorégressive, les GM-Estimateurs sont résistants en présence de "additive outliers" sous des conditions de régularité ainsi que pour une foncton $\eta(\cdot,\cdot)$ bornée et continue. Cependant, lorsque le processus possède une partie de moyenne mobile, l'estimation à l'aide des GM-Estimateurs n'est plus robuste qualitativement (MARTIN et YOHAI 1985 P.136).

Ces observations pourraient amener l'utilisateur à ne pas se servir des GM-Estimateurs. Cependant, la perte d'efficacité en présence de "innovation outliers" semble faible comparativement au gain apporté par la diminution du biais en présence de "additive outliers" (DENBY et MARTIN 1979). Ainsi, lorsque l'on ne connaît pas la structure exacte des valeurs douteuses, il peut être intéressant d'utiliser les GM-Estimateurs.

BIBLIOGRAPHIE

BOX G.E.P et G.M. JENKINS(1976). Time Series Analysis : Forecasting and Control, revised edition. Holden-Day, San Francisco.

BUSTOS A. et M. AHSANULLAH(1984). On the Estimates of the First Order Autoregressive Process. Commum. Statist.-Simula. Computa., 13(4), 489-505.

BUSTOS O.H.et V.J. YOHAI(1986). Robust Estimates of ARMA Models. JASA, vol. 81, No. 393, 155-168.

FOX A.J.(1972). Outliers in Time Series. Jour. Royal Stat. Society, ser B, 34, 350-363.

FULLER W.A.(1976). Introduction to Statistical Time Series. John Wiley & Sons, Inc. New York.

HAMPEL F.R. (1971). A General Qualitative Definition Of Robustness. Annals of Math. Stat. 42, no.6. 1887-1896.

HUBER P.J.(1981). Robust Statistics. John Wiley & Sons, New York. 308 p.

MARTIN R.D. et V.J. YOHAI(1985). Robustness in Time Series and Estimating ARMA Models.In: Handbook of Statistics 5,Ed. D.R. Brillinger and P.R. Krishnaiah. New York : Elsevier, 119-155.

MARTIN R.D.(1982). The Cramer-Rao Bound and Robust M-Estimates for Autoregressions. Biometrika 69,437-442.

MARTIN R.D.(1980). Robust Estimation in Autoregressive Models. In: D.R. Brillinger and G.C. Tiao, eds., Directions in Time Series, 228-254. Institute of Mathematical Statistics Publication, Haywood, CA. 228-254.

MARTIN R.D. et L. DENBY(1979). Robust Estimation on the First Autoregressive Parameter. JASA 74,140-146.

TUKEY J.W.(1976). Useable Resistant/Robust Techniques of Analysis. In: W.L. Nicholson et J.L. Harris, eds., Proc. First ERDA Statistics Symposium. Batelle Northwest Laboratories, Richland, WA.

APPLICATION DES SERIES CHRONOLOGIQUES A L'ETUDE DES REVISIONS

Bernard Lefrançois

Division des mesures et de l'analyse des industries
Statistique Canada
Ottawa, Ontario
K1A 0T6

Résumé: *Les révisions apportées à certaines composantes des Comptes nationaux canadiens sont étudiées comme séries chronologiques afin de déterminer s'il est possible, et comment il serait possible d'améliorer les données provisoires couramment publiées en les rapprochant des premières données révisées annuellement. De meilleurs estimés de la variation temporelle seraient ainsi obtenus. Les principaux résultats sont que les données provisoires peuvent potentiellement être améliorées et que cette amélioration découlerait d'une modification aux méthodes d'ajustement aux jalons.*

1. INTRODUCTION

De nombreuses données produites par la plupart des organismes statistiques sont d'abord publiées provisoirement afin de répondre au besoin des utilisateurs en données d'actualité, et sont par la suite révisées pour y inclure l'information qui n'était pas connue au moment de la publication initiale. Le sacrifice d'un peu de précision au bénéfice de l'actualité peut cependant créer des problèmes lorsque l'on veut faire des comparaisons inter-temporelles, puisqu'en règle générale les données concernant une période antérieure ont déjà été révisées alors que celles se rapportant à la période courante ne l'ont pas encore été. Si, par exemple, nous voulons mesurer la variation (en %) du Produit national brut (PNB) entre le premier et le second trimestre de 1987 au moment de la diffusion des données du second trimestre, la pratique actuelle est d'utiliser toute l'information disponible, soit les données provisoires du second trimestre et les données révisées du premier.

Or, cette mesure de la variation n'est utilisée que parce que nous supposons que les chiffres révisés sont de meilleurs estimateurs du niveau des séries que les chiffres non révisés, du fait qu'ils reposent sur plus d'information. Mais cela n'en fait pas pour autant la meilleure mesure de la variation. Rao, Srinath et Quenneville(1987) montrèrent qu'il faut évaluer les propriétés statistiques des données provisoires par rapport aux données révisées pour déterminer le meilleur estimateur de la variation, ce qui nous amène à l'étude des révisions proprement dites. Cet article est le résumé d'une étude, disponible sur demande, montrant comment l'analyse des séries chronologiques peut servir à l'étude des révisions apportées aux données produites périodiquement. Cette analyse nous permet de vérifier si les données provisoires peuvent être améliorées et dans quelle mesure. Elle nous permet aussi de déterminer s'il est possible d'améliorer l'estimation de la variation d'une période à l'autre. De plus, ce genre d'analyse nous renseigne abondamment sur la 'qualité' des données.

Dans la section 2 nous passons en revue les études antérieures sur les révisions et caractérisons les révisions optimales comme étant imprévisibles. Dans la section 3, les séries étudiées, soit les premières révisions annuelles apportées aux comptes nationaux du Canada, sont décrites et analysées sommairement. Ceci révèle que les révisions apportées à *16 des 24* principales composantes brutes des comptes nationaux, ainsi que celles apportées à *18 des 24* principales composantes dessaisonalisées, sont autocorrélées. Il est également observé que cette autocorrélation ne peut être attribuée à la dessaisonalisation mais que le niveau d'agrégation pourrait l'expliquer en partie. Dans la section 4, nous analysons en détail une de ces séries autocorrélées de révisions en reliant les données provisoires aux données révisées annuellement à l'aide d'un modèle de fonction réponse. Ce modèle montre que les données provisoires sont d'assez bonnes estimations des données révisées. Il y a, cependant, des indications à l'effet qu'il serait possible d'améliorer les estimations provisoires, en modifiant les méthodes d'ajustement aux jalons actuellement en usage.

2. REVISIONS OPTIMALES ET SERIES CHRONOLOGIQUES

De bonnes révisions doivent être petites par rapport à la précision d'une mesure. Mais la précision d'une mesure n'est pas toujours disponible, comme nous permettent de le constater les agrégats économiques, car ils sont généralement formés à partir de nombreuses sources d'information, certaines pour lesquelles il est difficile, sinon impossible, de contrôler la précision. Même lorsque les révisions sont jugées de peu d'importance, elles peuvent se comporter de façon systématique, ce qui peut fausser la trajectoire de la variable, un aspect très important de nombreuses mesures. Ainsi, quel que soit le mouvement systématique décelé dans les révisions, de quelque importance qu'elles soient, il devrait être corrigé à moins que la déformation de la trajectoire de la variable soit jugée négligeable. Suivant Mankiw, Runkle et Shapiro(1984), de bonnes données provisoires ne devraient pas être susceptibles d'aucune amélioration, si l'on tient compte de toute l'information disponible au moment de leur publication; une donnée provisoire améliorée étant plus près de la donnée révisée que la donnée non révisée. Conséquemment, de bonnes révisions devraient être imprévisibles en ce sens que la meilleure prévision d'une révision future devrait être nulle, étant donné l'information disponible au moment de la publication des données provisoires. Mais, contrairement à ces auteurs, nous n'utiliserons pour toute information que le comportement de séries chronologiques des données provisoires et révisées.

En définissant P_t comme l'estimation provisoire pour la période t, et F_t comme l'estimation révisée ultérieurement pour la même période, la révision (R_t) apportée à l'estimation provisoire s'exprime comme suit:

$$R_t = F_t - P_t$$

En termes statistiques, l'imprévisibilité des révisions optimales se traduit par une espérance nulle à toutes les périodes, et puisque la connaissance de la révision effectuée dans une période antérieure ne doit pas permettre de déterminer la valeur des révisions effectuées dans des périodes ultérieures, les révisions ne doivent pas être corrélées dans le temps, c'est-à-dire:

$$E(R_t) = 0 \text{ et}$$
$$Cov(R_t, R_{t-k}) = 0 \text{ pour toute valeur de t et pour tout } k \neq 0.$$

Dans le langage des séries chronologiques, un tel processus stochastique s'appelle un bruit blanc, et si les révisions sont effectivement des bruits blancs, on ne peut espérer améliorer les données provisoires puisque leur meilleure prédiction est zéro. On ne peut non plus espérer améliorer la méthode d'estimation de la variation actuellement en usage.

On distingue quatre grandes catégories de travaux dans la littérature économique et statistique portant sur les révisions et plus spécialement sur celles apportées à des agrégats économiques comme les comptes nationaux. Dans la première catégorie, les travaux visent à établir quelques-unes des propriétés statistiques de base des révisions, par exemple la moyenne, la variance, l'étendue, *etc.* Dans la deuxième catégorie, les études visent à déterminer l'effet des révisions sur les modèles économétriques, c'est-à-dire s'il y a changement, ou non, dans l'estimation de leurs paramètres. Les statistiques descriptives de base n'apportant que peu d'information sur les séries chronologiques, et puisqu'il n'existe pas encore de lois économiques quantitativement indéniables, aucune de ces deux approches ne pourrait nous être utile pour améliorer les données si jamais un aspect indésirable des révisions s'y manifestait.

Les études de la troisième catégorie analysent dans le détail chaque source d'erreur et de révision dans une mesure. Même si en théorie il est possible de le faire, cela peut être trop coûteux en pratique. De plus, compte tenu de l'éventail d'estimations produites par Statistique Canada, il est souhaitable que l'analyse des révisions ait un champ d'application très étendu. D'où la nécessité d'une approche globale. L'analyse des séries chronologiques est d'application étendue, elle est beaucoup moins coûteuse qu'une analyse détaillée de chaque source de révision, et elle permet d'identifier certains aspects indésirables dans les données.

Les études des révisions comme séries chronologiques peuvent être partagées en deux groupes. Dans le premier groupe on essait de déterminer si les révisions se comportent de la façon optimale décrite précédemment. Dans le second groupe, ce ne sont pas les révisions qui sont étudiées, mais la relation entre les données provisoires et les données révisées. Dans ces études, on considère les données provisoires comme des prédicteurs des données révisées. Cependant, la plupart des études recensées dans ce deuxième groupe firent dès le départ des hypothèses sur les valeurs déphasées des séries provisoire et révisée apparaissant dans les équations les reliant. C'est ce que nous appellerons la structure de la relation. Comme les causes de révision sont nombreuses (par exemple, réponses en retard, imputation, dessaisonalisation, ajustement aux jalons, biais dans les données provisoires attribuable à des mouvements saisonniers ou cycliques), il n'y a aucune raison de s'attendre à priori à une structure spécifique dans la relation entre les données révisées et les données provisoires. La structure de cette relation doit être vérifiée. C'est ce que nous avons fait avec quelques unes des séries des comptes nationaux du Canada, en utilisant les méthodes usuelles d'analyse des séries chronologiques.

3. LES REVISIONS DES COMPTES NATIONAUX CANADIENS.

Comme bien d'autres variables, les comptes nationaux des revenus et des dépenses du Canada sont révisés souvent, et ces révisions peuvent être classées selon leurs diverses fréquences. Il y a d'abord les révisions trimestrielles, effectuées dans les trimestres suivant celui de la publication initiale, et fondées sur des renseignements additionnels. Il y a ensuite les révisions annuelles, effectuées lors de la publication des données provisoires du premier trimestre d'une année, qui touchent les données des années précédentes. Ces révisions annuelles sont souvent, mais non uniquement, le résultat d'un ajustement aux jalons, c'est-à-dire l'incorporation dans une série d'information sur son comportement annuel. Ces renseignements ne sont jamais disponibles avant la fin d'une année, et plusieurs années sont révisées d'un seul coup. Ces années peuvent être révisées de nouveau lorsque s'ajoutent d'autres informations annuelles. Enfin, il y a aussi des révisions portant sur une période beaucoup plus longue, comme celles découlant du recensement.

Les séries étudiées sont énumérées dans le tableau 3.1. Les trois séries de révisions les plus intéressantes à analyser, par rapport aux données provisoires, sont les premières révisions trimestrielles, les premières révisions annuelles et les estimations les plus récentes, car ce sont les chiffres les plus fréquemment utilisés pour mesurer la variation d'une période à l'autre. Toutefois, nous n'avons étudié que les premières révisions annuelles apportées aux données provisoires en dollars courants, tant sous la forme brute que dessaisonalisée, s'étendant du premier trimestre de 1971 au dernier trimestre de 1984, parce qu'il n'était pas possible de créer une série de révisions homogène à partir des premières révisions trimestrielles, ni à partir des estimations les plus récentes.

La première partie de l'analyse était de nature descriptive et consistait principalement en des graphiques et de quelques statistiques sommaires de base. Elle avait pour but d'identifier les séries dont les premières révisions annuelles seraient les plus susceptibles d'améliorations. La figure 3.1 montre le graphique des premières révisions annuelles apportées à la série brute Rémunération des salariés. Ce graphique illustre bien les graphiques de la plupart des autres séries étudiées, brutes ou dessaisonalisées, où l'on constate que ces révisions ne sont pas aléatoires. Afin de déterminer avec exactitude le caractère non-aléatoire de ces révisions, nous avons calculé la statistique Q de Ljung-Box(1978) pour déterminer si la fonction d'autocorrélation (FAC) échantillonnale des révisions différait sensiblement de la FAC théorique d'un bruit blanc, égale à zéro à tous les décalages $k \neq 0$.

Suivant l'hypothèse que la FAC théorique est zéro, la statistique Q de Ljung-Box est asymptotiquement distribuée selon la loi du chi carré avec 24 degrés de liberté, soit le nombre de décalages de la FAC échantillonnale utilisés dans son calcul. Comme on peut le constater dans le tableau 3.1, la plupart des révisions sont autocorrélées, ce qui réfute l'hypothèse qu'elles soient optimales. En fait, seulement 8 des 24 séries brutes et seulement 6 des 24 séries dessaisonalisées peuvent être considérées comme non corrélées. D'après ce tableau, il est aussi évident que cette corrélation temporelle des révisions n'est pas uniquement transitoire ou illusoire: pour de nombreuses séries l'effet est très significatif.

FIGURE 3.1
PREMIERES REVISIONS ANNUELLES DE LA REMUNERATION BRUTE DES SALARIES

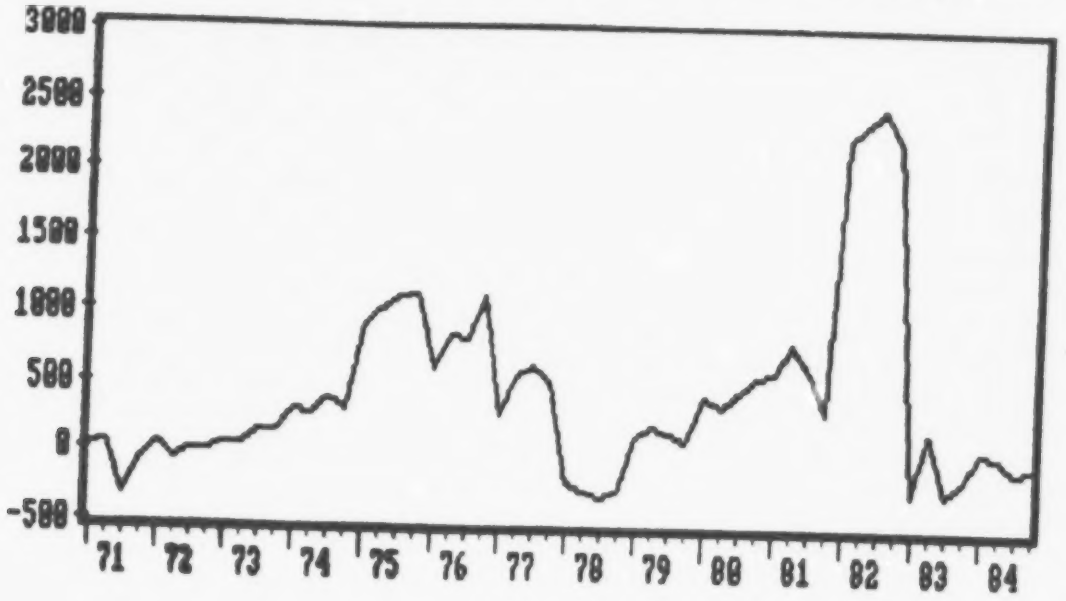


FIGURE 3.2
REGROUPEMENTS PAR DEGRE DE CORRELATION

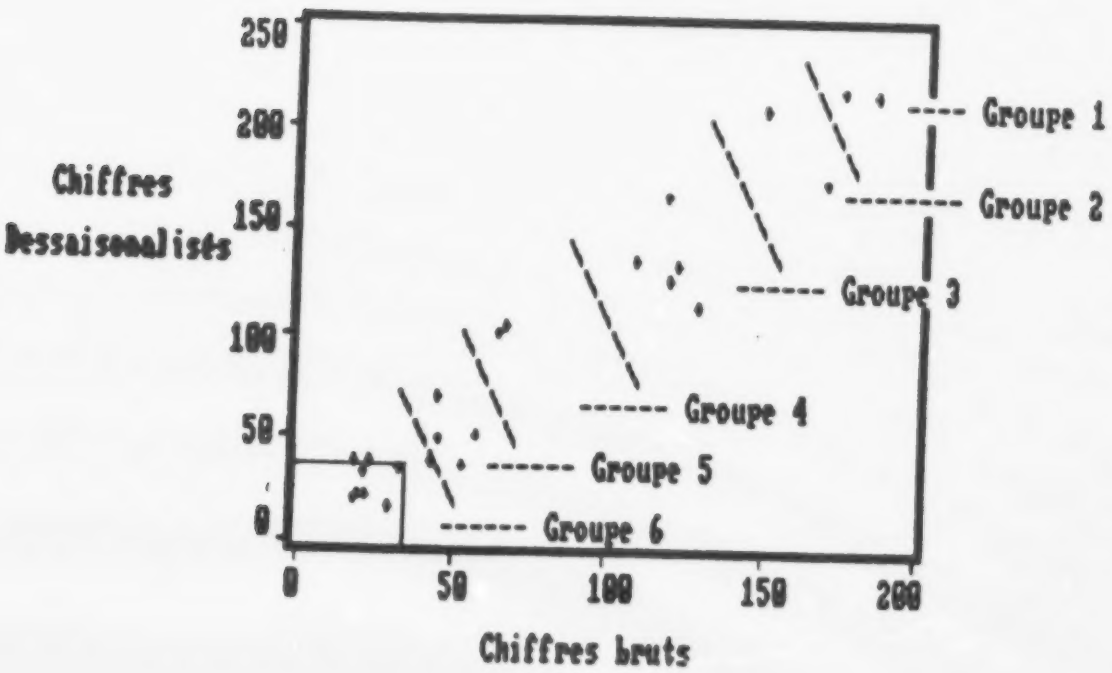


TABLEAU 3.1

SERIES DES COMPTES NATIONAUX DES REVENUS ET DES DEPENSES A L'ETUDE

NOM DE LA SERIE	GROUPE	STATISTIQUE DE LJUNG-BOX	
		Brute	Dess.
Comptes des revenus			
Rémunération des salariés	3	118.8***	126.4***
Solde et indemnités militaires	3	107.5***	136.2***
Bénéfice des sociétés			
- Dividendes versés aux non-résidents			
+ Revenus divers de placements			
- Ajustement de la valeur des stocks	5	43.8**	69.5**
Revenu comptable net des expl. agricoles	6	17.9	39.0*
Revenu net des entreprises indiv. non-agr.	4	63.4***	101.0***
Revenu national net au coût des facteurs	2	148.4***	210.1***
Impôts indirects moins subventions	4	66.0***	105.4***
Provisions pour consommation de capital	3	121.3***	133.5***
Erreur résiduelle	6	21.0	32.8
Produit national brut	1	173.6***	218.3***
PNB excl. Revenu comptable des expl. agr.	1	184.5***	216.5***
Comptes des dépenses			
Dépenses pers. en biens et services	3	117.4***	167.0***
Dépenses publiques en biens et services incl. la FBCF et la VVMS	3	128.1***	113.8***
Dépenses des entrep. en constr. résid.	6	23.0	38.7*
Dépenses des entrep. en constr. non-rés.	5	44.4**	49.9**
Dépenses des entrep. en mach. et matériel	5	56.4***	51.0**
Formation brute de capital fixe (FBCF) des entreprises	5	52.5***	37.5*
Valeur de la variation matérielle des stocks (VVMS) - entrep. non-agr.	6	18.9	22.6
VVMS - entreprises agricoles	6	18.1	20.2
VVMS - entreprises non gouvernementales	6	32.1	34.8
Exportations de biens et services	6	41.6*	37.8*
Importations de biens et services	6	21.4	21.9
Balance commerciale	6	29.2	16.2
Demande intérieure finales (incl. la FBCF et la VVMS publiques)	2	168.4***	174.8***

Légende :
* : significatif à un seuil de 95%
** : significatif à un seuil de 99%
*** : significatif à un seuil de 99.9%

Un aspect intéressant du tableau 3.1 est la similitude du résultat de l'application du test d'autocorrélation à une série brute de révisions et de son application à une série dessaisonalisée. La figure 3.2 est un nuage de points, chacun représentant les valeurs observées de la statistique Q pour une série de révisions sous ses deux formes, brute et dessaisonalisée. Une correspondance y est bien apparente. Cette correspondance dans les degrés de corrélation mesurés par la statistique Q pour les révisions apportées aux séries brutes et les révisions apportées aux séries dessaisonalisées nous porte à conclure que l'autocorrélation observée ne peut être attribuée à la dessaisonalisation. En nous servant de la figure 3.2, nous pouvons grouper les séries en fonction des valeurs décroissantes de la statistique Q. Ces groupes sont identifiés dans le tableau 3.1. Cette classification montre que la plupart des agrégats, comme le PNB et le PNB excluant le revenu net des exploitants agricoles, sont plus autocorrélés que leurs composantes, mais il y a des exceptions. Le niveau d'agrégation peut donc expliquer en partie cette autocorrélation. Nous avons donc concentré notre attention sur les séries du troisième groupe puisque leurs révisions sont fortement autocorrélées et puisque ces séries ne sont pas des agrégats des autres séries étudiées.

4. ANALYSE CHRONOLOGIQUE DE LA REMUNERATION DES SALARIES BRUTE.

La présente section est un compte rendu de l'analyse en séries chronologiques des premières révisions annuelles apportées à la série brute de la Rémunération des salariés, révisions qui furent identifiées comme fortement autocorrélées dans la section précédente. L'examen de la fonction d'autocorrélation (FAC) de la série des révisions montre clairement qu'elle n'est pas aléatoire, et indique que le principal facteur de ce caractère serait auto-régressif du premier ordre. Dès lors, nous pouvons conclure qu'il vaudra la peine d'identifier et d'estimer un modèle de fonction réponse (une relation linéaire temporelle) entre la série révisée annuellement pour la première fois et la série de données provisoires. En appliquant les méthodes habituelles d'identification et d'estimation des modèles de fonction réponse, décrites par Box et Jenkins(1976) et Gottman(1981), pour les observations s'étendant de 1971-T1 à 1984-T4, nous obtenons la relation suivante entre la série révisée annuellement (RSA_t) et la série provisoire (RSP_t), après les avoir transformées pour en stabiliser la variance au cours du temps:

$$\underset{}{log(RSA_t)} = \underset{[3807.1]}{1.00125}\ log(RSP_t) + \underset{[4.0]}{(1 - .56\ B^{10})}e_t/\underset{[5.8]}{(1 - .65\ B)}$$

Les e_t forment une suite de résidus non-corrélés, normalement distribués, d'espérance nulle et d'écart-type égal à .0112. Les chiffres entre crochets sont les valeurs de la statistique t des paramètres. Dans ce modèle, 'B' désigne l'opérateur de retard défini par $B\ X_t = X_{t-1}$.

Les principales caractéristiques de ce modèle sont les suivantes. Premièrement, il ne requiert pas de constante pour l'ensemble de la période étudiée, indiquant que les données provisoires constituent des estimations non 'biaisées' des données révisées annuellement. Deuxièmement, la 'pente' estimée est significativement supérieure à un (1.0). Cette sous-estimation des données révisées par les données provisoires est de l'ordre de 1.4%. Troisièmement, et comme on s'y attendait, ce modèle a une composante autorégressive du premier ordre qui supprime une bonne partie de

l'autocorrélation des révisions. Enfin, le terme d'erreur suit un mouvement saisonnier; mais il convient de souligner que ce paramètre de moyenne mobile n'est qu'approximatif parce qu'il n'élimine pas tout l'effet saisonnier observable, et parce qu'il est instable. Un changement de la période d'estimation modifierait sensiblement sa valeur et peut-être même son ordre, mais un quelconque paramètre saisonnier serait toujours requis.

Dans l'ensemble, les modèles des séries étudiées montrent que les chiffres provisoires semblent être d'assez bons prédicteurs des chiffres révisés annuellement pour la première fois, hormis l'autocorrélation des résidus. Pouvons-nous maintenant utiliser ces modèles pour améliorer les données provisoires publiées et, ainsi, l'estimation de la variation tel que suggéré par Conrad et Corrado(1979) et par Rao, Srinath et Quenneville(1987)? Car si nous pouvions utiliser ces modèles pour prévoir les futures données révisées annuellement en n'utilisant que les données provisoires, nous pourrions publier en premier ces prévisions au lieu des chiffres provisoires, ce qui réduirait l'ampleur des futures révisions annuelles. Les estimations publiées de la variation serait, *de facto*, moins sujettes à révision.

Malheureusement la réponse est non. Si, par exemple, nous réestimons le modèle de la Rémunération brute des salariés en utilisant les observations s'étendant de 1971-T2 à 1983-T4, nous aurions obtenu les estimations du tableau 4.1 pour l'année 1984. Ce tableau montre que les données prédites sont plus éloignées des données révisées que les données provisoires effectivement publiées. Ceci se produit parce qu'il y a un 'saut' dans la relation entre les données révisées annuellement et les données provisoires lorsque l'on passe du quatrième trimestre de 1983 au premier trimestre de 1984, ce saut n'étant pas prévu par le modèle. Puisque les révisions annuelles sont effectuées pour toute une année en même temps, cette information n'aurait pas été disponible au moment opportun pour les fins de la prévision.

TABLEAU 4.1
SERIE BRUTE DE LA REMUNERATION DES SALARIES

	Données provisoires publiées	Données révisées publiées	Données révisées prédites
1984-T1	55156	55185	55374
1984-T2	58076	58081	58552
1984-T3	59505	59395	60028
1984-T4	59627	59559	60016

Ce modèle, ainsi que les modèles des autres séries étudiées, n'améliore donc pas les données provisoires même si les tests d'adéquation habituels indiquent le contraire. Le 'saut' observé en passant du quatrième trimestre d'une année au premier trimestre de l'année suivante marque, tout au long des années observées, tous les modèles analysés puisque les résidus appréciables ont tendance à apparaître dans le premier trimestre d'une année. La répartition par trimestre de la somme des carrés des résidus du modèle de la Rémunération brute des salariés en fait foi, montrant que les résidus du premier trimestre représentent à eux seuls près de 70 % du total de la somme des carrés des résidus.

Donc, les résidus ne peuvent être considérés comme formant une suite de variables aléatoires identiquement distribuées, ce qui invalide toute inférence statistique usuelle reliée à ces modèles, en particulier l'évaluation de l'ampleur de la sous-estimation des données révisées par les données provisoires. Sommes-nous Gros-Jean comme devant ? Non, car l'incapacité des modèles d'améliorer les données provisoires, la concentration des résidus appréciables dans les premiers trimestres et l'instabilité de la composante saisonnière des termes d'erreur sont toutes des indications que les premières révisions annuelles sont largement attribuables à l'ajustement aux jalons.

L'ajustement aux jalons vise à incorporer des renseignements sur le comportement annuel d'une série chronologique observée infra-annuellement de sorte que le mouvement infra-annuel de la série ajustée soit le plus près possible de celui de la série non-ajustée. Les nombreuses méthodes actuellement en usage se distinguent par leur mesure de la distorsion du mouvement infra-annuel. Or, les diverses méthodes employées ne tiennent pas compte des données non encore observées, ce qui entraîne que les plus récentes données ajustées sont moins fiables que les plus anciennes parce que composées d'un élément aléatoire plus important dû à ce manque d'information. De plus, les jalons des années les plus récentes ne sont pas généralement disponibles, soit à cause du délai inévitable pour obtenir des estimés fiables, ou tout simplement parce que l'année visée n'est pas encore terminée. Il faut donc raccorder les plus récentes observations infra-annuelles aux observations ajustées, ce qui est généralement fait par l'application des taux de croissances des données non-ajustées à la dernière observation ajustée. Si cette dernière observation contient une forte composante aléatoire, les données raccordées la contiendront alors elles-aussi.

Un an après qu'un ajustement aux jalons ait été effectué, les données seront de nouveau ajustées en utilisant une période d'observation mise à jour d'une année. Un nouvel élément aléatoire sera alors introduit dans les données nouvellement ajustées. L'écart entre les chiffres provisoires de l'année précédente (composés de l'ancien élément aléatoire), et les chiffres de la même année révisés annuellement pour la première fois (composés du nouvel élément aléatoire) aura alors une composante égale à la différence de ces deux éléments aléatoires. Cette composante sera donc présente, et constante, dans chacune des observations de l'année en question. Les premières révisions annuelles auront donc une composante, aléatoire d'une année à l'autre, mais constante au court d'une même année. Le comportement de ces révisions sera donc similaire à celui des résidus des modèles de fonction réponse décrit précédemment. On comprend maintenant pourquoi ces modèles peuvent difficilement prévoir les futures données ajustées.

Est-il néanmoins possible d'améliorer les données provisoires avant leur publication ? La réponse est oui et les résultats obtenus nous montrent comment. Nous avons remarqué qu'avec la connaissance de l'observation ajustée du premier trimestre d'une année particulière, il était possible d'obtenir de bonnes prévisions des données ajustées des trois autres trimestres. Ainsi, à l'aide d'une quelconque information sur la valeur probable des jalons futurs, un modèle ARMMI par exemple, il serait possible d'améliorer les données provisoires d'une série qui sera ajustée à des jalons. Une méthode pour ce faire pourrait comprendre les étapes suivantes: premièrement, à partir de la série formée des jalons, prédire au moins celui pour l'année d'intérêt à l'aide d'un modèle ARMMI; ensuite, avec un modèle

ARMMI de la série provisoire, prévoir les valeurs non encore observées des autres périodes de l'année; enfin, ajuster la série provisoire augmentée de ses prévisions en utilisant la série de jalons elle aussi augmentée de ses prévisions.

On doit remarquer la similitude de cette approche avec l'amélioration apportée par Dagum(1975) à la méthode de dessaisonalisation X-11 du *U.S. Bureau of the Census*. Cholette(1984) arriva aussi à la conclusion que les jalons non encore observés devraient être prédits d'une quelconque façon. De même que la taille des révisions attribuables à la dessaisonalisation est réduite par la méthode améliorée X-11-ARMMI, la méthode décrite ci-dessus devrait réduire la taille des révisions attribuables à l'ajustement aux jalons, et, par conséquent, la taille des premières révisions annuelles.

REMERCIEMENTS

Je tiens à remercier Darryl Rhoades, Ruth McDougall et K.P. Srinath pour leurs commentaires et leurs conseils qui ont grandement amélioré cette étude.

BIBLIOGRAPHIE

Box, G.E.P., Jenkins, G.M.(1976) *Time Series Analysis: Forecasting and Control* Deuxième édition, Holden-Day, San Francisco, Californie.

Cholette, P.A.(1984) *L'ajustement des séries infra-annuelles aux repères annuels* Techniques d'enquêtes, *10*, 39-53

Conrad, W., Corrado, C.(1979) *Application of the Kalman Filter to Revisions in Monthly Retail Sales Estimates* J. of Econom. Dynamics Control, *1*, 177-198

Dagum, E.B.(1975) *Seasonal Factor Forecasts from ARIMA models* Présentation faite à la 40e session de l'Institut International de Statistique, Varsovie, Pologne.

Gottman, J.M.(1981) *Time Series Analysis: A Comprehensive Introduction for the Social Scientists* Cambridge University Press, Cambridge, Angleterre.

Ljung, G.M., Box, G.E.P.(1978) *On a Measure of Lack of Fit in Time Series Models* Biometrika, 65, 297-303

Mankiw, N.G., Runkle, D.E., Shapiro, M.D.(1984) *Are Preliminary Announcements of the Money Stock Rational Forecasts?* J. of Monetary Economics, *14*, 15-27

Rao, J.N.K., Srinath, K.P., Quenneville, B.(1987) *Optimal Estimation of Level and Change Using Current Preliminary Data* A paraître, Proc. Internat. Symp. on Panel Surveys (Washington, D.C., Nov. 19-22,1986)

ESTIMATION OPTIMALE DU NIVEAU ET DE LA TENDANCE EN UTILISANT LES DONNEES PRELIMINAIRES

J.N.K. Rao, K.P. Srinath et B. Quenneville
Université Carleton et Statistique Canada

RESUME

Il est coutimier dans les enquêtes-entreprises de publier un estimé préliminaire pour la période en cours et un estimé final pour la période précédente. La différence entre les deux estimés donne l'estimé de la tendance. Il est possible que cet estimé de la tendance soit biaisé et très différent de la différence entre les deux estimés finaux. Il est donc opportun d'étudier le problème d'obtenir un meilleur estimé préliminaire qui donnera, en conséquence, un meilleur estimé de la tendance. Deux approches sont considérées. La première, nommée approche de l'échantillonnage, utilise une moyenne pondérée de l'estimé préliminaire de la période en cours et de l'estimé préliminaire de la période précédente ajustée pour la tendance. La deuxième approche utilise les méthodes de séries chronologiques en modèlisant les estimés finaux ainsi que la différence entre les estimés finaux et préliminaires. Les résultats sont étendus pour inclure l'erreur d'échantillonnage dans les estimés.

REFERENCE

Rao,J.N.K.,Srinath,K.P. et Quenneville,B.(1986) 'Estimation of Level and Change Using Current Preliminary Data' Proceedings of the International Symposium on Panel Surveys, Washington, D.C., Wiley, (à paraitre).

UTILISATION DE FICHIERS ADMINISTRATIFS À DES FINS STATISTIQUES: ESSAI D'ANALYSE DE LA MOBILITÉ DES JEUNES TRAVAILLEURS

Danielle Bilodeau
Louise Bourque
Françoise Tarte
Bureau de la statistique du Québec

INTRODUCTION

Le fichier du registre des gains des cotisants de la Régie des rentes du Québec est utilisé comme fichier de base pour effectuer cette étude. Un échantillon de 39 560 travailleurs présents au fichier durant les années 1966 à 1981 a été tiré à partir de ce fichier. Un plan de sondage a été élaboré au préalable de façon à assurer la précision voulue et la représentativité de cet échantillon. La banque de données longitudinales ainsi constituée contient un historique des différents gains et employeurs pour chaque travailleur de l'échantillon. L'utilisation de tels fichiers administratifs à des fins statistiques soulève de nombreuses difficultés qui seront illustrées dans ce texte. Quelques résultats préliminaires sur la mobilité des jeunes travailleurs au Québec sont ensuite présentés.

1. HISTORIQUE DU PROJET

Il est bon de débuter par un historique du projet pour illustrer toutes les étapes et tous les délais nécessaires lorsqu'on veut effectuer une étude à partir de fichiers administratifs.

En avril 1982, la Régie des rentes du Québec et le Bureau de la statistique du Québec ont convenu d'évaluer la faisabilité d'une étude sur la mobilité des travailleurs. On a alors examiné la possibilité d'utiliser le fichier du registre des gains des cotisants de la Régie des Rentes comme fichier de base pour la réalisation de cette étude.

De mai 1982 à mai 1985, on a construit le fichier final qui servira à l'étude sur la mobilité des travailleurs. Le Bureau de la statistique du Québec a donné les spécifications à la Régie des rentes du Québec pour tirer un échantillon de travailleurs à partir du fichier du registre des gains. Des détails sur ces spécifications sont donnés à la section 2.

Pour réaliser l'étude, il a fallu compléter l'information contenue dans le fichier du registre des gains. Cette information est décrite plus en détail à la section 2. Le fichier du registre des gains contient des données de type socio-économiques telles que l'âge, le sexe, etc. sur les individus et le numéro de chaque entreprise pour lesquelles ils ont travaillé au cours des années 1966 à 1981. Il a fallu pairer ce fichier au fichier central des entreprises pour obtenir des informations sur les entreprises telles que le secteur industriel, etc.

Ce pairage des fichiers a été effectué au ministère du Revenu du Québec, car seul ce Ministère, pour des raisons de confidentialité, pouvait le faire légalement. De plus, le ministère du Revenu a effectué une reconversion des numéros d'entreprise inscrits au fichier du registre des gains en un numéro unique se rapportant à l'entreprise-mère. Cette dernière opération sera explicitée plus en détail à la section 2.

Les difficultés inhérentes à un tel projet sont nombreuses. L'étape de construction du fichier a duré environ 3 ans. Trois principaux facteurs peuvent expliquer ce délai: En effet, trois organismes étaient impliqués dans ce processus de construction et les opérations effectuées devaient l'être de façon séquentielle. De plus, des projets plus prioritaires pour les organismes impliqués ont parfois pris le pas sur le projet de construction du fichier. Concernant l'aspect confidentiel du projet, des délais ont aussi été occasionnés par des demandes d'autorisation pour effectuer le pairage des deux fichiers nécessaires à l'étude. Le fichier final ne contient que les numéros d'assurance sociale masqués des travailleurs préservant ainsi la confidentialité des informations contenues dans le fichier.

La construction du fichier a nécessité des ressources humaines et monétaires importantes de la part des trois organismes impliqués. Depuis juin 1985, le Bureau de la statistique du Québec et la Régie des rentes du Québec ont effectué différents tests de validation sur le fichier afin d'en éliminer les observations aberrantes, ou encore les artifices générés par le processus administratif. Par exemple, certains enregistrements ont dû être effacés, car ils étaient en fait des duplicata de gains déjà inscrits au fichier: on pouvait s'en rendre compte à l'aide d'une variable de type administratif définissant le code de transaction. Depuis juin 1985, ces deux organismes ont également effectué différents essais pour tenter de cerner le concept de la mobilité d'un travailleur à l'aide de différentes mesures.

La construction du fichier et sa préparation pour fins d'utilisation ont donc, comme le démontre l'historique qui précède, nécessité beaucoup d'efforts de la part de nombreuses personnes. Les résultats préliminaires obtenus à partir de ce fichier au début de 1987 et présentés ici à la section 5 sont toutefois assez encourageants et justifient, en l'absence de toute autre base de données longitudinales sur la mobilité des travailleurs, au Québec du moins, les efforts investis dans ce projet.

2. DESCRIPTION DU FICHIER[1]

Avant d'aborder la description du fichier proprement dite, définissons certains termes qui seront utilisés tout au long de ce texte:

travailleur:

travailleur, autonome ou salarié, inscrit au registre des gains des cotisants de la Régie des rentes du Québec (RRQ). Pour être inscrit au registre, un travailleur doit avoir versé des cotisations au RRQ. Celles-ci sont prélevées à partir du revenu de travail qui excède l'exemption de base de cotisation au régime des rentes du Québec pour l'année considérée. Cette cotisation est obligatoire pour un travailleur âgé entre 18 et 70 ans. (On entend par travailleur indifféremment un homme ou une femme).

entreprise:

entreprise-mère telle que définie dans le fichier central des entreprises. C'est la plus grande entité légale qui emploie un travailleur. Un travailleur peut changer d'établissement physique et demeurer toujours à l'emploi de la même entreprise-mère. Ce travailleur a été considéré comme n'étant pas mobile pour les fins d'étude

---

1. Le texte de Guy Nadeau et de Jean-François Boivin de la Régie des rentes du Québec intitulé "Mobilité des travailleurs: plan de travail" a servi dans une large mesure à la rédaction de la première partie de cette section.

de la RRQ. Une entreprise-mère peut avoir plusieurs numéros d'employeur, ces numéros étant liés à plusieurs systèmes de comptabilité, pour les fins de l'impôt. C'est pour cette raison qu'on a dû reconvertir les numéros d'employeur du fichier du registre des gains en un numéro unique d'entreprise-mère obtenu du fichier central des entreprises.

mobilité:

il y a mobilité d'un travailleur lorsqu'un travailleur change d'entreprise tel que défini plus haut. Si un travailleur change d'occupation, d'établissement ou de lieu de travail géographique mais qu'il demeure toujours à l'emploi de la même entreprise, il n'est pas considéré comme étant mobile.

période étudiée:

1966 à 1981

territoire couvert:

tout travail exercé, à titre d'autonome ou de salarié, sur le territoire québécois dans une entreprise faisant affaires au Québec.

hors-marché:

il y a absence de gains pour une année donnée pour les raisons suivantes:

- sortie du marché du travail
- chômage
- maladie
- décès
- congé sans solde
- gains inférieurs à l'exemption de base
- non résident du Québec au 31 décembre.
- ne fait pas partie de la population active.
- moins de 18 ans.
- à la retraite.

Il faut noter qu'on ne peut identifier dans le fichier les périodes de retrait du marché du travail, de chômage, de maladie ou de congé sans solde inférieures à une année.

Voici maintenant les grandes lignes du plan de sondage ayant servi à tirer un échantillon représentatif constituant le fichier construit pour l'étude sur la mobilité.[1]

---

1. La partie de cette section concernant le plan de sondage est tirée presqu'intégralement du texte de Louise Bourque du Bureau de la statistique du Québec intitulé "Enquête sur la mobilité des travailleurs: plan de sondage".

L'univers de l'étude est formé de l'ensemble des travailleurs du Québec actifs entre 1966 et 1981. La base de sondage est constituée des travailleurs du Québec inscrits au fichier du registre des gains des cotisants de la Régie des rentes du Québec entre 1966 et 1981. L'unité de base est le travailleur présent au fichier du registre des gains durant les années 1966 à 1981.

Afin de réduire le plus possible la taille de l'échantillon, on a stratifié l'univers selon deux variables. La première de ces variables est stable dans le temps: c'est le sexe du travailleur. La seconde par contre l'est moins, c'est le statut du travailleur. Toutefois, selon une étude menée sur un échantillon préliminaire, on constate une tendance de la population active à la constance au niveau du statut.

Comme la variable statut du travailleur est fonction du temps, on a choisi de stratifier l'univers selon le statut à un moment particulier: en 1966, en 1981 si hors-marché en 1966 et en hors-marché en 1966 et en 1981. Les années 1966 et 1981 ont été choisies comme critères de stratification afin que le plus de travailleurs possibles se retrouvent dans les strates autres que hors-marché en 1966 et en 1981. La variable statut d'emploi est formée de trois catégories: travailleur salarié, autonome ou mixte. Un travailleur a un statut de travailleur mixte lorsqu'il a occupé des emplois à titre d'autonome et à titre de salarié au cours de la même année. On a donc formé 14 strates qui sont explicitées au tableau 1.

On donne aussi au tableau 1 le nombre d'individus dans l'échantillon et dans la population pour chaque strate ainsi que le poids associé à chacune des strates. On peut voir dans ce tableau que l'échantillon compte 39 705 travailleurs et constitue 8% de la population constituée de 5 181 889 travailleurs.

Cette taille d'échantillon a été déterminée de façon à estimer un certain nombre de proportions selon l'année (1966 à 1981), le sexe et le groupe d'âge de façon relativement précise. Il est important de noter qu'il s'agit d'un échantillon d'individus et non pas d'entreprises. Le tirage de l'échantillon est un tirage systématique, les travailleurs étant classés au préalable par ordre croissant de numéro d'assurance-sociale. Pour chacun des 39 705 travailleurs de l'échantillon le fichier contient un enregistrement pour chaque emploi occupé pour chacune des années pendant lesquelles il a travaillé. Chaque enregistrement contient les informations suivantes:

<u>Informations provenant du registre des gains</u>:

- numéro d'assurance sociale masqué
- sexe
- date de naissance
- année d'imposition
- type de travail (autonome, salarié)
- numéro d'employeur
- code de transaction
- gains totaux

- contributions au Régime des rentes du Québec
- date de décès

Informations provenant du fichier central des entreprises:

- numéro d'entreprise
- code d'activité économique
- taille de l'entreprise
- autres variables relatives à l'entreprise.

Au total, le fichier contient environ 450 000 enregistrements. On se rappellera que le numéro d'entreprise unique utilisé pour mesurer la mobilité provient du fichier central des entreprises. Ce numéro est présent pour chacun des enregistrements. Par contre, les autres informations provenant du fichier central des entreprises seront difficilement utilisables car elles étaient absentes pour environ 40% des enregistrements.

En construisant ce fichier, nous avons obtenu une banque de données longitudinales issue de fichiers administratifs faisant le suivi de 39 705 travailleurs pendant 16 années qui couvrent la période 1966 à 1981.

Tableau 1

Strate	Échantillon	Population	Poids
Femmes autonomes en 1966	2,800	8,400	3
Femmes salariées en 1966	2,885	620,265	215
Femmes mixtes en 1966	598	1,794	3
Hommes autonomes en 1966	4,274	85,482	20
Hommes salariés en 1966	1,941	1,358,692	700
Hommes mixtes en 1966	1,516	37,898	25
Hors marché en 1966, femmes autonomes en 1981	1,757	19,333	11
Hors marché en 1966, femmes salariées en 1981	4,292	923,740	215
Hors marché en 1966, femmes mixtes en 1981	601	10,818	18
Hors marché en 1966, hommes autonomes en 1981	1,267	44,376	35
Hors marché en 1966, hommes salariés en 1981	1,459	1,024,756	702
Hors marché en 1966, hommes mixtes en 1981	534	26,691	50
Hors marché en 1966 et en 1981, femmes	11,275	565,099	50
Hors marché en 1966 et en 1981, hommes	4,506	454,545	101
TOTAL	39,705	5,181,889	

3. AVANTAGES ET INCONVÉNIENTS RELIÉS À L'UTILISATION DE FICHIERS ADMINISTRATIFS

3.1 Inconvénients

Voici quelques inconvénients liés à l'utilisation de fichiers administratifs auxquels nous avons eu nous aussi à faire face:

- Les variables présentes dans les fichiers sont souvent différentes de celles que l'on voudrait mesurer idéalement.

- Il est nécessaire de bien comprendre les processus administratifs sous-jacents à la constitution de ces fichiers pour pouvoir les "nettoyer" de certaines données "fausses", si l'on considère ce qui "devrait" se trouver dans le fichier, mais qui ont été tout de même enregistrées dans les fichiers pour des besoins administratifs.

- Les démarches nécessaires pour obtenir les fichiers nécessitent un grand nombre de ressources et occasionnent de nombreux délais. Les coûts associés à la construction d'un fichier pour fins d'étude sont donc assez importants. On peut toutefois se demander quels auraient été les coûts associés à une enquête ayant la même taille d'échantillon répétée pendant 16 années.

- Les plans de sondage ont souvent moins de flexibilité par rapport à une enquête étant données les contraintes imposées par un système déjà existant.

- Les fichiers offrent souvent une couverture partielle de la population à l'étude.

- Ces fichiers sont souvent sujets à de nombreux changements, soit au niveau des définitions des variables, soit au niveau de l'enregistrement de données dans le fichier. Par exemple, lorsque les spécifications pour le tirage de l'échantillon ont été données, les chiffres des populations de travailleurs dans les strates n'étaient plus les mêmes que ceux utilisés pour déterminer les tailles d'échantillon. Il a fallu ajuster la pondération en conséquence. De plus, certains travailleurs ont changé de strates lors du processus de tirage de l'échantillon. Il faut tenir compte de ces changements lors du traitement.

3.2 Avantages

Malgré ses nombreux inconvénients, l'utilisation de tels fichiers administratifs possède tout de même certains avantages par rapport à l'utilisation d'enquêtes ou de recensements:

- En général, la qualité des données monétaires nécessaires à l'administration est assez bonne. Souvent même, elle est meilleure que lorsqu'on pose la question à un répondant qui n'a pas nécessairement en mémoire ou sous la main le chiffre exact.

- Dans le fichier du registre des gains, l'âge et le sexe étaient présents à 99%.

- Il y a un meilleur suivi des travailleurs car il n'y a pas de perte due à la non-réponse dans le temps.

- Au Canada, il n'existe pas d'enquête permettant de mesurer la mobilité des travailleurs sur une période aussi longue que 16 années.

4. MESURE DE LA MOBILITÉ D'UN TRAVAILLEUR

Le concept de mobilité qu'on voudrait idéalement mesurer se définit ainsi: un travailleur qui change d'entreprise au cours d'une période donnée. Compte tenu de l'information disponible dans le fichier, on a finalement mesuré quelque chose de plus facile à observer: un travailleur qui ne change pas d'entreprise au cours d'une période donnée. Ce travailleur est alors défini comme stable.

On observe un travailleur à la période initiale t. On l'observe de nouveau à la période suivante t + 1. Si le travailleur occupe le même emploi au cours de ces deux périodes, on dit qu'il est stable. Dans notre cas, les périodes d'observation sont des périodes annuelles.

Cette mesure génère toutefois deux problèmes losqu'on la superpose aux données du fichier:

- pour les travailleurs autonomes occupant un ou plusieurs emplois au cours d'une année donnée.

- pour les salariés occupant plusieurs emplois au cours d'une année donnée.

Les solutions suivantes ont été adoptées pour intégrer ces cas problématiques dans la définition de mobilité donnée ci-dessus:

- Les travailleurs autonomes. Un travailleur autonome est considéré comme étant lui-même une entreprise. S'il a eu plusieurs emplois en tant qu'autonome au cours d'une année, on additionne ces gains et on considère qu'il ne s'agit que d'une seule entreprise.

- Les salariés à emplois multiples. Le fichier contient un très grand nombre de travailleurs occupant plusieurs emplois au cours d'une année donnée. Or, la seule information présente dans le fichier concerne les gains pour chaque emploi mais on ne connait

pas la durée de ces emplois, la date de début de l'emploi et la date de fin de l'emploi. On ne peut donc pas savoir, pour une année donnée, si le travailleur occupe ces emplois de façon consécutive ou concuremment. De plus, il est difficile de déterminer, lorsqu'un travailleur occupe deux ou trois emplois à chaque année, s'il a réellement changé d'emploi en comparant deux années consécutives. Dans la mesure de mobilité, on tient compte finalement de plus d'un emploi. Si le travailleur a conservé les mêmes emplois pendant deux années consécutives, il est stable. On ne considère toutefois que 4 emplois principaux au maximum.

Dans un premier temps, une mesure simplifiée et opérationnelle de type "stable ou non stable" pour une année donnée semble la plus appropriée. Dans un deuxième temps, on pourra faire des tests de sensibilité en modifiant cette mesure pour tenir compte de diverses considérations. On pourrait, par exemple, étudier plus en profondeur les travailleurs à emplois multiples.

Plus précisément, on définit la stabilité au temps t + 1 pour chacun des cas suivants:
Cas 1

On retient l'entreprise ayant généré le plus de revenus à chaque année. Si l'entreprise au temps t + 1 est la même que celle au temps t, l'individu est défini comme stable au temps t + 1.

Cas 2

Si le travailleur n'est pas déjà classé comme stable et qu'il a travaillé pour au moins deux entreprises au temps t et au temps t + 1, on retient les deux entreprises ayant généré le plus de revenus à chaque année. Si ces deux entreprises sont les mêmes au temps t et au temps t + 1, l'individu est défini comme stable au temps t + 1.

Cas 3

Si le travailleur n'est pas déjà classé comme stable et qu'il a travaillé pour au moins trois entreprises au temps t et au temps t + 1, on retient les trois entreprises ayant généré le plus de revenus à chaque année. Si ces trois entreprises sont les mêmes au temps t et au temps t + 1, l'individu est défini comme stable au temps t + 1.

Cas 4

Si le travailleur n'est pas déjà classé comme stable et qu'il a travaillé pour quatre entreprises au temps t et au temps t + 1, on retient les quatre entreprises ayant généré le plus de revenus à chaque année. Si ces entreprises sont les mêmes au temps t et au temps t + 1, l'individu sera défini comme stable au temps t + 1.

La stabilité est ainsi définie pour les années 1967 à 1981. Un travailleur non stable est défini comme étant mobile.

Cette mesure de mobilité possède certains avantages:

- Avec cette mesure, les travailleurs saisonniers occupant plusieurs emplois à chaque année ont la possibilité d'être considérés comme stables.

- On néglige les emplois passagers et peu importants, comme par exemple de petits contrats occasionnels, ce qui simplifie de beaucoup l'analyse des données.

- On peut faire des comparaisons avec les grandes enquêtes annuelles qui utilisent souvent le concept d'emploi principal dans l'année ou de dernier emploi occupé.

5. RÉSULTATS PRÉLIMINAIRES ET COMPARAISON AVEC D'AUTRES ÉTUDES

Les résultats qui suivent portent sur une partie de l'échantillon seulement, soit les jeunes travailleurs. Les travailleurs âgés entre 18 et 29 ans pour chacune des années de 1966 à 1981 ont été sélectionnés. Au total, 21 113 travailleurs ont été sélectionnés. Pour faciliter le traitement informatique, la banque de données a dû être transformée. Au départ, on avait un enregistrement par année par emploi par travailleur. On l'a modifié pour obtenir un enregistrement par travailleur en ne gardant que quatre emplois principaux pour une année. Cette information peut sembler un peu technique mais le grand nombre d'informations contenues dans la banque de données demandaient des efforts au niveau de la programmation pour éviter que les coûts informatiques ne deviennent astronomiques. La banque de données se prête aussi plus facilement de cette façon à des études de type longitudinal.

On présente au tableau 2 le pourcentage de travailleurs stables pour les années 1967 à 1981. Ces résultats sont détaillés par sexe. Il s'agit d'une stabilité annuelle en ce sens que, par exemple, si un travailleur en 1967 occupait le ou les mêmes emplois qu'en 1966, il est considéré comme stable en 1967.

Tous les tableaux présentés sont pondérés étant donné que l'échantillon est stratifié et que la fraction de sondage diffère selon les strates. Les résultats se rapportent donc bien à la population des travailleurs. Le pourcentage de travailleurs mobiles est égal à 100 moins le pourcentage de travailleurs stables.

Tableau 2
Pourcentage de travailleurs stables selon l'année et le sexe

Année	Hommes	Femmes	Total
1967/1966	63,9	72,2	66,8
1968/1967	69,0	71,7	70,0
1969/1968	65,2	72,6	67,9
1970/1969	70,0	74,0	71,5
1971/1970	66,4	71,6	68,4
1972/1971	65,5	68,7	66,8
1973/1972	63,9	68,7	65,8
1974/1973	61,0	67,2	63,5
1975/1974	62,1	66,4	63,9
1976/1975	63,8	67,6	65,3
1977/1976	51,7	60,0	55,0
1978/1977	61,6	65,0	63,0
1979/1978	60,0	65,6	62,3
1980/1979	61,2	64,0	62,4
1981/1980	60,0	67,0	62,8
Ensemble	62,6	67,5	64,5

Pour que la variable de stabilité soit définie pour une année donnée, il faut que le travailleur ait travaillé au cours de cette année et au cours de l'année précédente. (Tous nos calculs sont donc basés sur une période d'observation de durée de travail relativement longue). Ceci étant dit, on peut voir au tableau 2, que les femmes semblent plus stables que les hommes. Ainsi, losqu'on observe des travailleurs des deux sexes qui travaillent de façon régulière, les femmes ont moins tendance à changer d'entreprises que les hommes.

En ce qui concerne l'évolution de l'indice de stabilité au cours des années, l'indice diminue dans les périodes d'expansion économique et augmente dans les périodes de récession. Les points de retournement sont présentés au tableau 3. En période d'expansion, les travailleurs ont donc plus tendance à changer d'entreprises; les perspectives d'emploi étant meilleures, les changements d'emploi offrent un meilleur rendement. En période de récession, on observe le phénomène inverse, les travailleurs s'efforçant plutôt de conserver l'emploi qu'ils occupent à ce moment là.

On présente au tableau 4 les pourcentages de travailleurs stables selon l'âge pour l'ensemble des années 1966 à 1981. Dans ce tableau, on présente aussi les résultats publiés par Osterman (1980). Il a utilisé le "National Longitudinal Survey", une enquête longitudinale menée aux État-Unis, pour réaliser son étude. Sa définition diffère de la nôtre: un travailleur est stable si l'emploi occupé à l'année t est dans la même firme que l'emploi occupé à l'année t-1

ou dans la même industrie (codifiée à 3 chiffres). Il est tout de même intéressant de faire un parallèle entre les deux études car les deux permettent d'observer que les travailleurs, très instables autour de la vingtaine, sont de plus en plus stables à mesure qu'ils approchent de la trentaine. Ces résultats demandent à être fouillés plus en profondeur et il sera intéressant de les comparer aux indices de stabilité des travailleurs ayant plus de trente ans, ce qui sera la prochaine étape de l'étude sur la mobilité des travailleurs.

Deux autres auteurs, Byrne (1975) et Haber (1983), ont calculé des taux de mobilité et des taux de départ à partir de l'enquête américaine mensuelle sur l'emploi, le "Current Population Survey". Ces résultats sont présentés au tableau 5 et ils sont comparables à ceux que nous avons obtenus.

Tableau 3
Cycle de croissance au Québec[1]

Pic (P) ou Creux (C)	Datation des points de retournement du PIB commercial non agricole
C	1er trimestre 1968
P	1er trimestre 1969
C	1er trimestre 1971
P	1er trimestre 1974
C	2e trimestre 1975
P	1er trimestre 1977
C	1er trimestre 1978
P	2e trimestre 1981

1. Compilation spéciale, Bureau de la statistique du Québec, mai 1987.
2. Les points de retournement font référence à des phases successives de croissance forte et de croissance faible ou négative.

Tableau 4
Pourcentage de travailleurs stables selon l'âge

Âge	Nos résultats entre	Osterman
	1967 et 1981	1970
18 ans	47,3	48,4
19 ans	51,5	58,2
20 ans	55,0	70,7
21 ans	58,8	70,0
22 ans	62,1	70,1
23 ans	64,7	75,5
24 ans	68,2	77,0
25 ans	68,7	77,7
26 ans	72,4	80,2
27 ans	73,2	76,8
28 ans	75,8	80,3
29 ans	77,1	-
Ensemble	64,9	-

Tableau 5

Groupes d'âge	Byrne		Haber	
	Taux de mobilité		Taux de départ	
	1973		1977	
	Hommes	Femmes	Hommes	Femmes
18 à 19 ans	37,3	36,2	48,4	44,0
20 à 24 ans	24,9	18,7	34,7	33,0
25 à 34 ans	12,4	9,9	18,6	20,2

Osterman (1980) a aussi calculé des taux de départ à partir du National Longitudinal Survey: 65,0% pour les travailleurs âgés de 18 ans et 29,0% pour ceux âgés de 27 ans. Ici encore, on observe que les travailleurs, très instables autour de la vingtaine, se stabilisent à l'approche de la trentaine.

Les résultats obtenus à partir des grandes enquêtes américaines sont du même ordre de grandeur que ceux que l'on obtient avec le fichier du registre des gains. De plus, les ventilations selon le sexe et l'âge amènent aux mêmes conclusions. Les résultats préliminaires obtenus à partir du fichier du registre des gains sont donc très encourageants.

CONCLUSION

On a vu dans cet exposé qu'il est possible de construire des bases de données représentatives de la population à l'aide d'un plan de sondage. On a vu également qu'on pouvait définir des concepts de mobilité qui, sans correspondre exactement à la définition idéale de la mobilité que l'on retrouve dans la littérature, nous permettent d'obtenir une mesure de la mobilité des travailleurs.

L'utilisation de fichiers administratifs a causé de nombreux problèmes mais on a pu finalement obtenir des résultats intéressants à partir des ces fichiers. Le comportement observé chez les jeunes travailleurs québécois et chez les femmes du Québec est comparable à celui décrit dans la littérature sur ce sujet. De plus, ces résultats ayant trait à la mobilité des travailleurs québécois étaient inexistants auparavant.

BIBLIOGRAPHIE

Osterman, Paul, Getting Started: the Youth Labor Market, The MIT Press, 1980

Bourque, Louise, Enquête sur la mobilité des travailleurs - plan de sondage, Direction de la consultation et des méthodes, B.S.Q., Avril 1984.

Nadeau, G. et J.F. Boivin, Mobilité des travailleurs: plan de travail, version préliminaire, Mai 1984

Byrne, James J., "Occupational Mobility of Workers", Monthly Labor Review, Février 1975.

Haber, Sheldon E., Enrique J. Lamas et Gordon Green, "A New Method for Estimating Job Separations by Sex and Race", Monthly Labor Review, Juin 1983.

Utilisation d'un fichier administratif à des fins statistiques: l'exemple du fichier des allocations familiales du Québec.

par
Raymond Ouellette [1]

texte présenté au 55e congrès de l'ACFAS à Ottawa
le 21 mai 1987 dans le cadre du colloque sur les
applications de la statistiques

[1] Ministère de l'Education du Québec, Direction des études économiques et démographiques, 1035 rue de la Chevrotière, 8e étage, Québec, Qc G1R 5A5

Utilisation d'un fichier administratif à des fins statistiques: l'exemple du fichier des allocations familiales du Québec.

HISTORIQUE

Le recensement scolaire a été longtemps la source privilégiée d'information sur les populations scolaires et préscolaires. Mais cette source était de qualité fort inégale. Le fichier des allocations familiales, par contre, renferme des informations fiables, ne serait-ce qu'à cause de son aspect financier. Le régime est universel... pour le moment!

En 1979, les CS qui connaissaient les codes postaux qui formaient leur territoire, en envoyaient la liste à la RRQC qui, en retour, transmettait sous la forme d'une liste de codes postaux le nombre d'enfants âgés de 0 à 18 ans résidant sur leur territoire selon la langue de correspondance avec la Régie. Le code postal, un élément de codification géographique au découpage très fin en milieu urbain, est la principale clé de tri des données. La RRQC devait satisfaire un nombre de plus en plus important de demandes et elle considérait que ce rôle ne devait pas être joué par elle.

En avril de la même année, le MEQ et la RRQC conviennent d'un protocole d'accès et d'utilisation du fichier des allocations familiales. Les deux parties partagent un même souci de confidentialité des données. Le MEQ de son côté répondra au volume croissant des demandes d'informations des CS mais uniquement dans le cadre des projets d'immobilisations.

Durant l'été 1979 le MEQ (à la DEED) a établi des tables de codes postaux par CS, on a alors réparti géographiquement les 152 000 codes du territoire québécois pour les CS catholiques et pour les CS protestantes. Nos tables sont cumulatives et aucun code n'est retiré même s'il cesse d'exister. En effet, pour reconstituer rétrospectivement certains territoires, il faut tenir compte des anciens codes; c'est pourquoi notre fichier contient aujourd'hui plus de 153 000 codes!

Dès 1979 la DEED utilise le fichier des allocations familiales comme source de données du nombre d'enfants d'âge préscolaire pour ses prévisions de l'ensemble du Québec à la maternelle. Durant l'été 1980 les tables de correspondance de codes postaux des CS commencent à être utilisées et, en janvier 1981, le fichier des AF devient la source exclusive du nombre d'enfants de 0 à 18 ans au Québec pour les prévisions à la maternelle ventilées non seulement par CS mais aussi par secteurs de CS.

DESCRIPTION DU FICHIER

Le fichier des AF est un fichier administratif mais il contient des informations que nous pouvons ici utiliser à des fins statistiques. Tout ce qui concerne la gestion et le contrôle de l'émission des chèques ne nous intéresse pas. La copie du fichier que la RRQC nous envoie est volumineuse: il y a 1 enregistrement par famille avec structure variable, cela occupe 4 rubans magnétiques. La DEED traite ce fichier et le transforme: 1 enfant par enregistrement et une structure fixe. Il n'occupe plus qu'un seul ruban.

Voici les principales variables utilisées:

Adresse incluant le code postal: Le code postal est l'élément de codification géographique principal. Il permet de déterminer le nombre d'enfants qui résident sur un territoire si ce territoire est lui-même décrit sous la forme d'une liste de codes postaux. L'adresse complète sert dans les cas suivants: correction d'un code postal erronné, ajout d'un code postal absent, localisation plus précise des enfants dans le cas où un code postal couvre un territoire trop vaste (comme c'est souvent le cas en milieu rural) ou lorsqu'un code est partagé entre 2 ou plusieurs territoires scolaires.

Année et mois de naissance de l'enfant: Permet le calcul de l'âge à la fin de n'importe quel mois. L'entrée dans le système scolaire dépend de l'âge et les modèles démographiques de prévisions d'effectifs scolaires sont fondés sur cette caractéristique.

Sexe de l'enfant: Permet le calcul de taux de fréquentation scolaire selon le sexe. Le modèle informatisé de relevés de population par CS utilise cette variable.

Langue de correspondance: C'est la langue utilisée entre le bénéficiaire et la Régie (**F** français **A** anglais **E** indéterminé). Nos modèles de prévisions utilisent cette variable pour prévoir les entrées selon la langue d'enseignement. Cette variable sert également à l'étude des mouvements migratoires des enfants selon la langue (effets de la loi 101)

Statut mensuel (les 20 mois au complet): Cette variable technique permet, pour chacun des mois, de savoir s'il faut compter un enfant dans le cadre d'un relevé de population. Elle permet aussi de compter les enfants qui ont quitté le Québec, ce qui est utile pour la planification scolaire.

Raison de terminaison du versement de l'allocation familiale: Complément au statut mensuel, cette variable permet de distinguer les différents type de mouvements migratoires des enfants.

Code de province: Pour repérer les enfants qui ont un dossier sous la juridiction du Québec. Le fichier contient, à cause de son caractère cumulatif, des dossiers qui ne sont plus sous la juridiction québécoise.

Numéro du dossier: C'est la seule clé disponible entre les diverses versions du fichier au MEQ et le fichier original de la Régie. Ce numéro est attribué par la RRQC conjointement avec le Ministère des affaires sociales du Canada.

Suffixe de l'enregistrement: C'est la clé de définition de l'enregistrement, il permet de distinguer les différents types.

Nombre d'enfants dans le dossier: Cela permet de faire la cohérence lors de l'exploitation de l'enregistrement qui est de longueur variable. C'est également une réponse directe à la question "nombre d'enfants dans la famille".

L'enregistrement dont le numéro de dossier est 999999999 et le suffixe est 9: Cet enregistrement contient des données résumant le contenu de tout le fichier. Il permet de vérifier la justesse de nos exploitations. On y retrouve:

1. le nombre d'enfants selon 9 états différents comme: les actifs, en suspend/étude, transfert autre province, transfert autre compte/famille, les 18 ans, etc.
2. le nombre de dossiers selon 6 statuts possibles.
3. les dates de production du fichier.

Il faut noter que le jour de la naissance de l'enfant n'est disponible dans aucun fichier de la Régie. De plus le nom des parents et de l enfant ne sont jamais transmis par cette dernière. Le MEQ peut demander une version du fichier pas plus de 3 fois par année à la RRQC, en pratique nous demandons la version d'octobre (assimilée au 30 septembre) et celle d'avril (6 mois plus tard). Le MEQ s'est engagé à diffuser ces renseignements uniquement sous la forme de tableaux statistiques.

Le MEQ possède une version annuelle du fichier depuis celui de 1974. De 1980 à aujourd'hui nous en avons 2 versions par année (octobre et avril). On distingue 2 sortes de fichiers AF-MEQ: un fichier dit historique et un fichier dynamique. Dans un fichier dynamique, le code postal de résidence du bénéficiaire est le dernier à jour; c'est le cas de nos fichier depuis 1979. Mais dans un fichier historique le code postal correspond à l'adresse du bénéficiaire 1 1/2 année plus tard à cause des alléas du système. Par programmation la version DEED a replacé ces codes à la bonne année (position) pour les fichiers de 1974 à 1977.

fichiers historiques	fichiers dynamiques
1974	juillet 1979
1975	janvier 1980
1976	avril 1980 à 1987
1977	octobre 1980 à 1986

APPLICATIONS ET UTILISATIONS DU FICHIER DES ALLOCATIONS FAMILIALES

Nous produisons des documents statistiques, des études et analyses ainsi que des prévisions scolaires en exploitant ce fichier.

1. **Documents statistiques:** Il s'agit essentiellement de deux relevés de la population âgée de 0 à 18 ans au 30 septembre par territoire de commission scolaire.
   Le premier selon le code postal en ordre alpha-numérique, il s'agit d'une production d'environ 5 000 pages pour l'ensemble du Québec. Chaque direction régionale du Ministère reçoit la partie qui lui est propre sous forme de listage informatique.
   Le second document est un relevé selon la confessionnalité, chaque CS connaissant ainsi sa population catholique ou protestante âgée de 5 à 18 ans (il n'y a pas de distinction chez les 0-4 ans) pour l'ensemble de leur territoire. Ce dernier document sert entre autre pour répartir la taxation scolaire.

2. **Etudes et analyses:** Nous produisons à partir des fichiers d'avril les nombres de départs du Québec vers d'autres pays ou vers d'autres provinces par âge, mois de l'année et langue de correspondance. Nos données débutent en 1974. Ces données n'ont été utilisées chez nous que pour étudier l'émigration des anglophones québécois (Maheu Robert dans Cahier québecois de démographie, vol. 12 no 2, octobre 1983).

3. **Prévisions démographiques:** Le fichier des allocations familiales permet un relevé de la population âgée de 0 à 4 ans fort précis, à jour et surtout selon un découpage géographique qui peut être très fin (en milieu urbain seulement). Ces données servent pour nos trois types de prévisions: la prévision de la revue de programmes de l'ensemble du Québec pour la gestion du Ministère, les prévisions automatisées par CS toujours pour la gestion du Ministere, et enfin les prévisions pour fins d'immobilisation scolaire, désagrégées à un niveau inférieur à celui de la CS.

LIMITES DES DONNEES

A cause de l'aspect financier pour les bénéficiaires, le fichier des allocations familiales est fort complet. Il est de plus cumulatif et mis à jour périodiquement.

En avril 1987 notre ruban dénombrait 2 562 621 personnes; je dis "personnes" et non "enfants" car un individu demeure inscrit au fichier aussi longtemps que le dernier membre de sa famille y ayant

droit n'en est pas exclu. Parmis tous nos inscrits, 95 413 n'ont pas de code postal et 52 584 ont un code hors du Québec. Un grand nombre des codes "à blanc" seraient des personnes résidant hors du Canada. Au Québec, les codes postaux commençant par G, H et J regroupent 94,2% des inscriptions, 3,7% des inscrits n'ont pas de codes et les 2,1% restant résident hors du territoire québécois. Le code postal est notre principale clé de tri et, fort heureusement, son usage est très généralisé.

Pour nos besoins de prévisions scolaires ce fichier est nettement supérieur aux défunts recensements scolaires. A peu près tout ceux qui y ont droit sont inscrits au fichier des AF, mais il y a des délais d'inscription. C'est surtout apparent chez les enfants de 0 an. En effet, on peut supposer que tous les enfants plus vieux qui sont nouvellement inscrits le sont avec un délai similaire quel que soit leur âge (ce sont des enfants nouvellement arrivés au Québec pour la plupart). Mais à 0 an la principale arrivée est, bien sûr, la naissance. Légalement les parents ont 4 mois pour inscrire leurs enfants mais il n'y a aucune obligation à demander les allocations familiales, seul l'aspect financier sert d'incitatif. A la Régie des rentes du Québec, on s'en remet au fichier fédéral pour toute modification à l'inscription, ajout, retrait, etc. Au fédéral les inscriptions permettent un paiement rétroactif d'au maximum 1 an.

Nos données d'enfants âgés de 0 an démontrent qu'il y a sous-estimation à cet âge et c'est surtout le délai d'inscription qui en est la cause. En prennant pour hypothèse que 100% des nouvelles naissances seront inscrites au fichier, alors le sous-dénombrement depuis 5 ans correspond environ à 11,6 % chez les francophones et 18,6 % chez les anglophones. Avec 95% de certitude, le délai d'inscription varie de 37 à 47 jours chez les francophones et de 58 à 76 jours chez les anglophones.

Lorsque nous faisons des prévisions du nombre d'enfants à la maternelle, nous utilisons les 5 dernières années disponibles des fichiers du 30 septembre de la RRQC. Pour chaque cohorte nous examinons l'accroissement absolu entre 2 années successives et nous retenons un accroissement (positif ou négatif) qui sera projeté. Comme les 0 an sont toujours sous-dénombrés et qu'il n'y a pas lieu de croire que ce sous-dénombrement varie grandement d'une année à l'autre, nous retenons alors des accroissements plus élevés pour le passage de 0 à 1 an ce qui corrige cette situation.

LE FICHIER DES AF A-T-IL AMELIORE LA JUSTESSE DE NOS PREVISIONS?

Le fichier des AF ne sert qu'à prévoir les élèves de la maternelle seulement. Evidemment l'impact, lorsqu'il y a erreur, peut être grand car la maternelle est l'entrée dans le système scolaire et une erreur s'y répercutera longtemps.

Les prévisions pour la revue de programmes ont été contre-vérifiées en janvier dernier. Les écarts entre la maternelle prévue et

la réalité varient entre 0,1% et 0,7% (dont 0,3% pour les 3 dernières prévisions) pour la première année et n'excèdent pas 1,4% avant 5 ans. Une seule prévision n'utilisant pas le fichier des AF a été retracée, elle se servait du nombre de naissances et couvrait la période 1973 à 1977: les écarts montaient très rapidement au dessus de 3%.

Evidemment ces prévisions impliquent de gros chiffres et les risques d'erreur sont plus faibles sur 90 000 enfants à la maternelle que sur les 150 d'un secteur d'une CS. Mais nous avons également vérifié la justesse de nos prévisions pour fin d'immobilisation scolaires. Pour ce faire, nous avons contre-vérifié deux groupes de prévisions: un premier groupe réalisé entre 1978 et 1980 et un second groupe réalisé entre 1983 et 1985. Dans les 2 cas il y a plus de 200 observations.

Dès la première année de prévision, l'écart entre le nombre d'enfants prévus à la maternelle et la réalité était de 12,3% en 1978-80 et de 6,5% en 1983-85. Ce qui différencie les deux groupes témoins c'est qu'en 1978-80 seulement 6% des dossiers utilisaient le fichier des allocations familiales (l'écart était de 5,4% pour ceux-ci) contre la totalité en 1983-85. Depuis l'utilisation du fichier des AF les écarts à la maternelle sont, dès la première année de prévision, inférieurs à 8,5% pour des nombre d'élèves plus petits que 100 et à moins de 4% dès que le nombre d'enfants impliqués excèdent 200. Avant l'utilisation de ce fichier, on obtenait une erreur inférieure à 5% lorsque les effectifs dépassaient 1 000.

CONCLUSION

Le fichier des allocations familiales du Québec est une source d'information fiable et fort utile aux limites peu contraignantes. Cette source est un cadeau statistique pour nous et il ne faudrait pas la perdre surtout si l'universalité des AF était mis en cause. Existe-t-il un autre fichier de rechange, une source de données équivalente? Nous pensons peut-etre au fichier de la RAMQ (Régie de l'assurance maladie du Québec) même si son nombre de bénéficiaires inscrits excède la population québécoise!

Par contre à la RAMQ le pointage des enfants ne peut pas avoir la même fréquence (mensuelle) qu'au fichier des AF, la précision d'un recensement à une date donnée s'en trouverait affectée. On ne peut au MEQ que profiter du fichier des AF ... tant qu'il dure!

EXPLOITATION DES FICHIERS ADMINISTRATIFS
A DES FINS STATISTIQUES

Regroupement des données sur les conducteurs
(permis, infractions et accidents)

Michel Pampalon
Direction de la statistique
Vice-présidence à la planification et
à la promotion de la sécurité routière
Régie de l'assurance automobile du Québec

Mai 1987

INTRODUCTION

La Régie de l'assurance automobile du Québec de par ses opérations relatives à la gestion du Code de la sécurité routière et à l'indemnisation des victimes de dommages corporels accumule journalièrement une foule de données. Ces données sont conservées dans des fichiers administratifs lesquels ne sont pas nécessairement construits pour faciliter les exploitations de nature statistique.

La Direction de la statistique a toutefois entreprise depuis plusieurs années une organisation des données les plus souvent demandées de façon à faciliter la production de statistiques. C'est le sujet que l'on traitera dans la première partie de cette présentation. Dans la deuxième partie, on traitera de certains problèmes de contrôle reliés à l'exploitation de fichiers administratifs. La troisième partie portera sur une application qui a consisté à réaliser un fichier regroupant les statistiques concernant le bilan routier des conducteurs en terme d'accidents et d'infractions commises.

I. LE SYSTEME D'INFORMATION STATISTIQUES

Le système d'informations statistiques, qui sert à répondre à la majorité des demandes de renseignements est composé de six fichiers.

Ces fichiers sont construits à partir des informations contenues dans les fichiers administratifs. Le traitement des fichiers administratifs exige des ressources informatiques importantes compte tenu du nombre d'observations et de variables qui y sont enregistrées. Afin de réduire ce temps nécessaire au traitement des données et aussi de faciliter la programmation servant à interroger les fichiers, plusieurs travaux ont été réalisés afin d'identifier, d'organiser et de structurer les données les plus souvent demandées ou utilisées dans les productions régulières.

Les six fichiers qui composent présentement le "système d'informations statistiques" peuvent être traités indépendamment l'un de l'autre.

Toutefois, compte tenu des besoins, plus d'un fichier peut être interroger à la fois. Chacun des fichiers contient une ou plusieurs variables qui permettent de relier les fichiers entre eux. Les principales sont le numéro de permis de conduire (titulaire), le numéro de plaque d'immatriculation (véhicule) et le numéro d'événement (accident).

La Direction de la statistique a aussi préparé ou obtenu de d'autres organismes certains fichiers afin de faciliter la production de certains résultats. C'est ainsi qu'il a été intégré au système d'information statistique, une table qui transforme le code postal en un code de municipalité. Cette table a été réalisée à partir des informations disponibles à la Société

canadienne des postes (Répertoire des codes postaux) et au Bureau de la statistique du Québec (Fichier du code géographique du Québec et Perspectives démographiques régionales et infrarégionales).

De plus, l'ensemble de ces travaux a été réalisé à l'aide du logiciel "SAS" et les fichiers sont aussi des fichiers "SAS".

1. Le fichier "Immatriculation"

Ce fichier contient les données nécessaires à la production des statistiques sur le nombre de véhicules en circulation. Son contenu permet de présenter le profil du parc automobile: usage, marque, modèle, année de fabrication, masse nette, cyclindrée et nombre d'essieux pour les camions (depuis 1987). Il contient aussi quelques renseignements sur les propriétaires: sexe, âge et lieu de résidence.

2. Le fichier "Permis de conduire"

Ce fichier regroupe les informations portant sur les titulaires d'un permis de conduire régulier. On y retrouve les caractéristiques des permis émis: privilèges accordés et date d'obtention du premier permis pour obtenir l'expérience de conduite. Les informations disponibles sur les titulaires concernent le sexe, l'âge et le lieu de résidence.

3. Le fichier "Rapports d'accidents"

Le fichier "Rapports d'accidents" contient les renseignements recueillis à l'aide du formulaire "Rapport d'accident" utilisé par tous les corps policiers du Québec lorsqu'ils sont appelés sur les lieux d'un accident. La Régie saisit informatiquement

tous les formulaires que lui font parvenir les corps policiers. Elle fait également une certaine validation des données transmises. Malgré cette validation, certaines informations obtenues sont difficilement utilisables (ceinture de sécurité, état apparent des conducteurs, état des véhicules, état apparent des piétons).

Ce fichier contient une foule de renseignements sur chacun des sujets suivants: moment de l'accident, lieu de l'accident, type d'accident, environnements et circonstances. De plus, il permet d'identifier les conducteurs et les véhicules impliqués, ainsi que les victimes qui ont subi des blessures.

4. Fichier "Indemnisation"

Le fichier "Indemnisation" permet de produire des statistiques sur les victimes pour lesquelles la Régie a versé les indemnités prévues dans la Loi sur l'indemnisation des victimes de dommages corporels sans égard à la responsabilité de quiconque.

Pour les demandes ouvrant droit à une rente, le fichier contient plusieurs renseignements sur la situation socio-économique des victimes (type d'occupation, état civil, nombre de personne à charge, revenu brut et net). On y retrouve également des renseignements sur le dossier médical des victimes (type de déficit anatomo-physiologique et durée d'invalidité obtenue en comptabilisant le nombre de jours où une rente a été versée).

Pour les autres demandes, les renseignements disponibles se résument au sexe, l'âge, le lieu de résidence et le type d'indemnité qui a été versé. Au moyen du numéro d'événement, on peut obtenir les caractéristiques des accidents dans lesquels ces victimes ont été impliqués.

5. Fichier "Infractions-Code de la sécurité routière"

Ce fichier permet d'enregistrer les infractions commises en vertu du Code de la sécurité routière et entraînant la perte de points d'inaptitude. Lorsqu'une personne accumule 12 points à l'intérieur d'une période de deux ans, son permis de conduire est révoqué.

Ce fichier contient les informations sur la nature de l'infraction, le nombre de points perdus, la date de l'infraction et la date de condamnation.

6. Fichier "Révocation/suspension"

Ce fichier permet d'identifier les infractions commises en vertu du Code criminel. Il permet aussi d'identifier les personnes qui ont subi une révocation ou une suspension de leur permis de conduire ou de leur immatriculation suite à l'atteinte de 12 points d'inaptitude ou au contrôle de l'assurance obligatoire. Egalement, la conduite d'un véhicule non assuré impliqué dans un accident ou la conduite d'un véhicule lorsque le permis est révoqué entraîne également la suspension du permis ou du droit d'en obtenir un.

Les informations contenues sur ce fichier sont le type de révocation ou suspension, la nature de l'infraction (Code criminel), la date de l'infraction, la date de condamnation, la date d'entrée en vigueur et la durée de la révocation ou de la suspension.

II. PROBLEMES RELIES A L'EXPLOITATION DES FICHIERS ADMINISTRATIFS

Un des objectifs recherchés dans l'élaboration du système d'informations statistiques est de s'assurer que d'une année à l'autre, les changements qui se produisent dans les données administratives n'affectent pas d'une façon importante les concepts statistiques et par voie de conséquence les résultats.

Dans certains cas, les efforts qui sont alloués à cette tâche sont importants. Par exemple, les statistiques que nous avons produites sur le nombre de véhicules en circulation pour l'année 1986 indiquaient une augmentation de près de 6% alors qu'on aurait dû s'attendre à une augmentation de 2% à 3,0%. Après avoir fait plusieurs vérifications pour s'assurer qu'il ne s'agissait pas d'une erreur de programmation, d'un changement de définition ou d'une modification dans la façon de conserver les données, nous sommes venus à la conclusion que cette augmentation était due à l'informatisation d'un plus grand nombre de bureaux émetteurs. Ce changement a augmenté la vitesse à laquelle sont enregistrés les véhicules et réduit le nombre de transactions rejetées compte tenu que les corrections peuvent être effectuées au moment du rejet. Les renseignements sur le parc automobile pour l'année 1986 seront donc accompagnés d'une note faisant état de cette situation.

Nous essayons également de porter une attention particulière à la documentation des variables qui composent le système d'informations statistiques. Un changement dans la définition d'une variable peut avoir des effets importants sur les résultats produits. Cette tâche est également fort exigeante.

Aussi, nous nous efforçons de conseiller les divers clients du système d'information statistique sur la meilleure façon d'aborder un problème ou de traiter une variable donnée. Dans les

fichiers administratifs, certaines variables sont de meilleure qualité que d'autres. Habituellement, les variables à caractère monétaire sont sujettes à une validation plus stricte que les variables à caractère statistique. Pour certaines variables, il est donc préférable de travailler à partir d'un échantillonnage afin d'intégrer dans les étapes du projet, une étape de validation. C'est dans cette perspective que nous essayons de répondre aux divers besoins des usagers lorsque possible.

III. LE DOSSIER DE CONDUITE DES CONDUCTEURS (INFRACTIONS ET ACCIDENTS)

Les statistiques présentées dans cette partie ont été obtenues en mettant en relation les informations enregistrées sur les fichiers "Permis de conduire", "Rapports d'accidents", "Infractions-Code de la sécurité routière" et "Infractions-Code criminel" et "Révocation/suspension".

Les résultats de ces travaux ont été enregistrés sur un fichier lequel comprend pour chaque conducteur, le nombre d'infractions (Codes de la sécurité routière et criminel) commises pour chacune des années 1982 à 1986 et le nombre total d'accidents (corporels et matériels) pour la même période. De plus, ce fichier contient les variables suivantes: sexe, date de naissance, date d'expiration du permis, date d'obtention du premier permis (expérience de conduite) et la municipalité de résidence pour chacune des années concernées.

Pour la période de 1982 à 1986, près de 2 millions de conducteurs ont commis au moins une infraction ou été impliqués dans au moins un accident.

Tableau 1: Nombre de conducteurs selon le type de dossier - 1982-1986

Type de dossier	Nombre	%
- Accidents et infractions	628 713	31,5
- Infractions seulement	863 359	43,3
- Accidents seulement	504 024	25,3
- Total	1 996 096	100,0

De ce nombre, 32% ont commis au moins une infraction et ont été impliqués dans au moins un accident; 43% ont commis des infractions seulement et 25% ont été impliqués dans des accidents seulement.

Sur une base annuelle, le nombre de conducteurs ayant commis des infractions ou ayant été impliqués dans des accidents représente environ 20% du nombre de titulaires de permis de conduire.

Tableau 2: Nombre de titulaires de permis de conduire et de conducteurs - 1982-1986

Année	Titulaires de permis de conduire	Conducteurs	
	Nombre	Nombre	Taux/1000 titulaires
1982	3 456 765	655 550	190
1983	3 521 520	662 785	188
1984	3 598 262	635 384	177
1985	3 687 021	710 514	193
1986	3 774 935	754 894	200
Moyenne annuelle	**3 607 701**	**683 825**	**190**

Tableau 3: Répartition procentuelle des conducteurs et taux pour 1 000 titulaires selon le type de dossier - 1982-1986

Année	Type de dossier					
	Infractions et accidents		Infractions seulement		Accidents seulement	
	%	Taux/ 1000 tit.	%	Taux/ 1000 tit.	%	Taux/ 1000 tit.
1982	10,5	20	62,8	119	26,6	51
1983	11,0	21	60,0	113	29,0	55
1984	11,2	20	53,0	94	35,8	63
1985	12,2	23	53,7	104	34,2	66
1986	11,5	23	57,1	114	31,4	63
Moyenne annuelle	**11,3**	**21**	**57,3**	**109**	**31,4**	**60**

Annuellement, 2% environ des titulaires commettent au moins une infraction et sont impliqués dans au moins un accident; 11% commettent que des infractions et 6% sont impliqués dans des accidents seulement.

On notera qu'au cours de la période, le taux pour 1 000 titulaires a augmenté particulièrement en 1985 et 1986 et diminué en 1984. On se rappellera qu'en 1984, en période de négociation, les policiers ont donné moins d'infractions. Pour les autres années, le nombre d'accidents survenus au Québec a augmenté à toutes les années depuis 1982, sauf pour l'année 1986 où l'on constate une légère diminution. Le nombre d'infractions a aussi augmenté depuis 1984.

La répartition des données selon le sexe indique que le taux pour 1 000 titulaires de sexe masculin est 2,2 fois plus élevé que celui des femmes. Ces taux ne tiennent toutefois pas compte du kilométrage parcouru par les conducteurs.

Tableau 5: Répartition procentuelle des titulaires et des conducteurs et taux pour 1 000 titulaires selon le sexe - moyenne annuelle 1982 à 1986

Sexe	Titulaires	Infractions		Accidents		Ensemble	
	%	%	Taux/ 1000 tit.	%	Taux/ 1000 tit.	%	Taux/ 1000 tit.
Masculin	58,5	80,6	179	76,2	105	77,9	225
Féminin	41,5	19,4	61	23,8	47	22,1	101
Total	100,0	100,0	130	100,0	81	100,0	190

Lorsque l'on considère que les infractions, le taux pour les conducteurs de sexe masculin est 2,9 fois plus élevé que celui des femmes alors que si l'on considère que les accidents, le taux obtenu est 2,2 fois plus élevé que celui des femmes. De plus, on a constaté que 21% des conducteurs de sexe masculin commettent

plus d'une infraction en moyenne annuellement alors que cette proportion pour les femmes est de 10%. En ce qui concerne les accidents, 11% des hommes sont impliqués en moyenne annuellement dans plus d'un accident alors que pour les femmes, cette proportion est de 5%.

Selon l'âge des conducteurs, ce sont les groupes entre 16-24 ans qui présentent le pire bilan: 25% environ des titulaires de ces groupes ont commis une infraction ou ont été impliqués dans un accident.

Tableau 6: Répartition procentuelle des titulaires et des conducteurs et taux pour 1 000 titulaires selon l'âge - moyenne annuelle 1982 à 1986

	Titulaires	Infractions		Accidents		Ensemble	
Age	%	%	Taux/ 1000 tit.	%	taux/ 1000 tit.	%	Taux/ 1000 tit.
16-19	5,4	7,0	169	7,6	115	6,9	242
20-24	13,4	18,8	182	18,0	109	17,8	253
25-34	28,4	32,5	149	29,7	85	31,4	210
35-44	22,3	22,3	130	20,5	74	22,0	187
45-54	14,6	11,4	101	12,3	68	12,0	156
55-64	10,3	6,0	75	7,9	61	6,9	127
65 et +	5,6	2,1	48	4,0	57	2,9	98
Total	**100,0**	**100,0**	**130**	**100,0**	**81**	**100,0**	**190**

Le seul autre groupe qui présente un taux supérieur à 200 est celui des 25-34 ans (210). Le groupe présentant le taux le plus faible est celui des 65 ans et plus (98). Ce résultat pour les personnes âgées est certainement relié en partie à une moins grande utilisation de leurs véhicules et par conséquent du nombre de kilomètres parcourus.

Concernant l'expérience de conduite, le taux de conducteurs pour 1 000 titulaires présente peu de variation pour les groupes dont l'expérience se situe entre 2 et 7 ans, soit environ 215 conducteurs pour 1 000 titulaires.

Tableau 7: Répartition procentuelle des titulaires et des conducteurs et taux pour 1 000 titulaires selon l'expérience de conduite - moyenne annuelle 1982 à 1986

Expérience de conduite	Titulaires	Infractions		Accidents		Ensemble	
	%	%	Taux/ 1000 tit.	%	Taux/ 1000 tit.	%	Taux/ 1000 tit.
1 an et -	7,3	6,7	118	8,4	93	7,0	182
2-3	7,9	9,1	151	9,5	98	8,9	215
4-5	8,3	9,7	154	9,5	94	9,4	217
6-7	8,5	9,8	149	9,1	87	9,4	210
8-9	8,7	9,7	143	9,0	82	9,4	202
10 ans et +	59,3	55,1	121	54,6	74	55,9	178
Total	**100,0**	**100,0**	**130**	**100,0**	**81**	**100,0**	**190**

Pour les autres groupes, les taux obtenus sont inférieurs. Il est à noter que pour le groupe de conducteurs de 1 an et moins, le taux de conducteurs ayant commis des infractions est le plus faible alors que celui concernant les accidents est l'un des plus élevés.

CONCLUSION

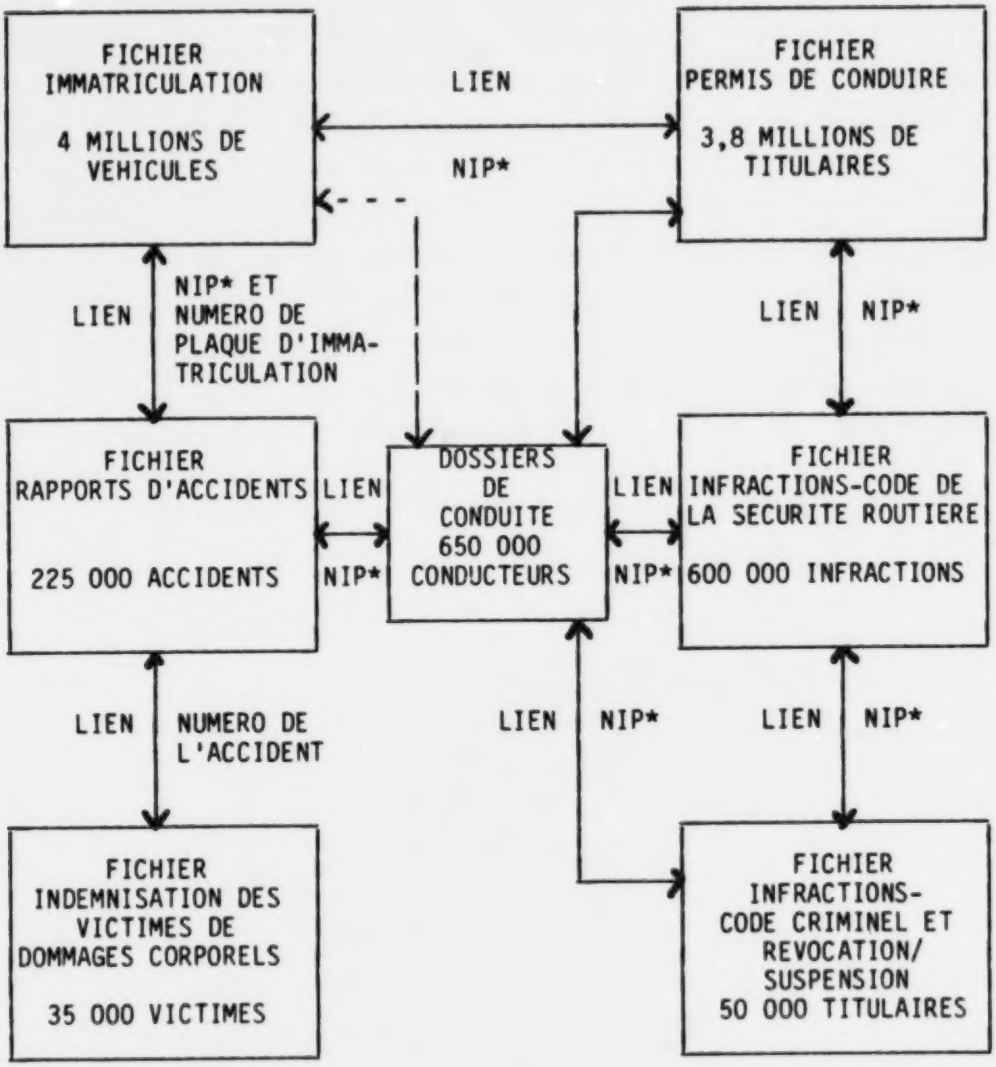
Le traitement des fichiers administratifs exigent des efforts importants tant au niveau de l'exploitation que du suivi de leur évolution. A la Direction de la statistique, nous avons choisi de se créer des fichiers abrégés à caractère statistique pour réduire les délais de réponse et faciliter les interrogations.

Egalement, pour faciliter les travaux qui exigent des traitements sur plusieurs fichiers, nous sommes à élaborer des fichiers de type "pivot" qui permettent de passer d'un fichier à un autre le plus simplement possible tout en réduisant le temps informatique nécessaire. C'est le cas du fichier "Dossier de conduite des conducteurs" utilisé pour produire les résultats de la Section 3.

A N N E X E

DESCRIPTION DU SYSTEME D'INFORMATIONS STATISTIQUES

ELEMENTS DU SYSTEME D'INFORMATIONS STATISTIQUES



* Pour les personnes physiques, il s'agit du numéro de permis de conduire.

FICHIER "IMMATRICULATION"

Autorisation

- Numéro de la plaque;
- Statut et période de validité.

Véhicule

- NIV: numéro d'identification du véhicule;
- Marque, modèle et année de fabrication;
- Masse nette, cylindrée et type de carburant;
- Masse totale en charge (avant 1987);
- Nombre d'essieux (depuis 1987).

Propriétaire

- NIP[1] : numéro d'identification du propriétaire;
- Type de personne (physique, corporation, société);
- Type propriété (locataire, propriétaire);
- Sexe et âge;
- Lieu de résidence (code postal, municipalité).

---

1. Pour les personnes physiques, ce numéro correspond au numéro de permis de conduire.

FICHIER "PERMIS DE CONDUIRE"

Autorisation

- Date d'expiration.

Permis de conduire

- Classes ou privilèges (autobus, véhicule de 11 000 kg et plus, taxi, véhicule de moins de 11 000 kg et motocyclette);
- Date d'obtention du premier permis (expérience de conduite).

Titulaire

- NIP: numéro d'identification du titulaire (ce numéro correspond également au numéro de permis de conduire);
- Sexe et âge;
- Lieu de résidence (code postal, municipalité).

FICHIER "RAPPORTS D'ACCIDENTS"

Identification

- Numéro de l'événement.

Moment

- Date de l'accident;
- Jour et heure.

Lieu de l'accident

- Municipalité et route;
- Intersection;
- Mercator (Sûreté du Québec seulement).

Type d'accident

- Gravité et nombre de victimes
- Genre d'accident;
- Nombre de véhicules impliqués.

Environnement

- Environnement et localisation;
- Catégorie de route et vitesse autorisée;
- Caractéristiques de la chaussée;
- Température et éclairage;
- Signalisation et visibilité.

Conducteur

- NIP: numéro de permis de conduire;
- Sexe, âge et expérience de conduite;
- Action et état du conducteur.

Véhicule

- Numéro de plaque et type de véhicule;
- Masse, année de fabrication, marque;
- Mouvement des véhicules.

Victimes

- Fonction et état de la victime (décès, blessures graves ou légères);
- Sexe, âge et port de la ceinture de sécurité.

FICHIER "INDEMNISATION"

Identification

- Numéro de la réclamation;
- Numéro de l'événement (accident).

Victime

- Sexe, âge, lieu de résidence;
- Type d'indemnité.

Rente (pour les demandes qui donne droit à une rente)

- Type d'occupation;
- Revenu brut et net;
- Rente journalière;
- Durée (nombre de jours où la Régie a versé une rente);
- Type de déficit anatomo-physiologique;
- Etat civil;
- Nombre de personne à charge.

FICHIER "INFRACTIONS-CODE DE LA SECURITE ROUTIERE"

Titulaire

- NIP: Numéro de permis de conduire.

Infraction

- Nature et date de l'infraction;
- Nombre de points d'inaptitude obtenus
- - Date de condamnation.

FICHIER "REVOCATION/SUSPENSION"

Titulaire

- NIP: Numéro de permis de conduire.

Révocation/suspension

- Type de révocation/suspension (Code criminel, atteinte de 12 points d'inaptitude, conduite durant révocation ou suspension, accident sans assurance, contrôle de l'assurance obligatoire);
- Nature de l'infraction (pour les articles du Code criminel);
- Date de l'infraction;
- Date de condamnation;
- Durée de la révocation/suspension.

LA MESURE DES ATTITUDES ET DES COMPORTEMENTS DANS LES ENQUÊTES PAR QUESTIONNAIRES*

François Béland**, Jean Lambert***, Brigitte Maheux**

Les recherches sur les attitudes et les comportements ont conclu à la nécessité de les mesurer au même niveau de spécificité et de généralité; soit vis-à-vis le même acte ou le même objet[1-13]. Pour mesurer attitudes et comportements envers le même objet ou le même acte, des questions à réponse fermées sont habituellement formulées de manière identique. Dans ce contexte, des biais systématiques de mesure apparaissent lorsque les personnes échantillonnées répondent d'une manière conséquente aux questions mesurant les attitudes et à celles mesurant les comportements à cause de leur proximité dans le même questionnaire.

Utilisant des données empiriques, la présence de biais systématiques causés par des réponses conséquentes à des questions formulées de même manière et mesurant attitudes et comportements, a été estimée. Les effets des biais systématiques ont été observés: 1) sur les moyennes et les variances des réponses aux questions relatives aux attitudes et aux comportements, 2) sur les différences entre les réponses aux questions d'attitudes et de comportements et 3) sur les corrélations entre attitudes et comportements envers le même acte ou objet.

DONNÉES ET MÉTHODES D'ANALYSE

Une étude sur les attitudes et les comportements des omnipraticiens dans leur rapport clinique aux patients a été conduite auprès d'un échantillon de 900 omnipraticiens. Les omnipraticiens échantillonnés ont été distribués au hasard en deux groupes. Le premier groupe a reçu un questionnaire comprenant les questions relatives aux attitudes et aux comportements. Le second groupe a reçu deux questionnaires, le premier mesurant les comportements et le deuxième les attitudes.

Le taux de réponse du sous-échantillon recevant un seul questionnaire a été de 64,3%, tandis que le taux de réponse sur le premier questionnaire pour le second groupe a été de 73,2%. Approximativement 10% des médecins répondant au premier questionnaire se sont abstenus de répondre au deuxième.

---

* Cette recherche a été rendue possible grâce à une subvention du Programme national de recherche et de développement en santé et par une bourse de chercheur attribuée aux deux premiers auteurs, par le Fonds de la recherche en santé du Québec.

** Groupe de recherche interdisciplinaire en santé, Faculté de médecine, Université de Montréal

*** Département de médecine sociale et préventive, Faculté de médecine, Université de Montréal

Les deux sous-échantillons ont été comparés sur les dimensions suivantes: l'âge, le sexe, la formation des médecins, l'école de médecine fréquentée, le type de pratique, le lieu de pratique et neuf caractéristiques qui décrivent le temps alloué à diverses activités professionnelles. Les échantillons n'étaient significativement différents sur aucune de ces dimensions.

Quarante et une questions relatives aux attitudes et quarante et une relatives aux comportements ont été présentées aux omnipraticiens. Les questions relatives aux attitudes et aux comportements ont été présentées de la même manière pour permettre de mesurer les attitudes et les comportements vis-à-vis le même acte ou le même objet.

Une analyse factorielle des réponses aux items d'attitudes a été effectuée sur chacun des sous-échantillons. Des 41 questions, 21 ont survécu aux tests de validité et de fiabilité. Les résultats de ces analyses factorielles ont été les mêmes dans chacun des sous-échantillons. De la même manière, l'analyse factorielle des réponses aux questions de comportements a donné le même résultat. Quatre dimensions ont été identifiées à l'aide de l'analyse factorielle: deux d'entre elles mesurent les rôles éducatifs du médecin (EDU.1 et EDU.2) alors qu'une troisième définit les aspects psychosociaux des soins médicaux (PSYCHO) et une dernière définit le contenu informatif de la pratique clinique du médecin (INF). Les coefficients de fiabilité des quatre dimensions sont tous au-dessus de .75 (voir en annexe les questions).

La cohérence des réponses aux questions d'attitudes et de comportements a été étudiée en observant 1) si les moyennes et les variances des distributions de réponses aux questions relatives aux attitudes (A) et aux comportements (C) différaient entre les sous-échantillons, 2) si les soustractions des réponses aux questions A de celles aux questions C correspondantes donnaient des résultats semblables ou différents selon les sous-échantillons et 3) si les corrélations entre A et C étaient significativement différentes entre les sous-échantillons. Des tests T ont été utilisés pour comparer les moyennes entre les sous-échantillons. Les coefficients de corrélation ont été transformés en scores-Z, puis un test de signification de leur différence a été appliqué[14].

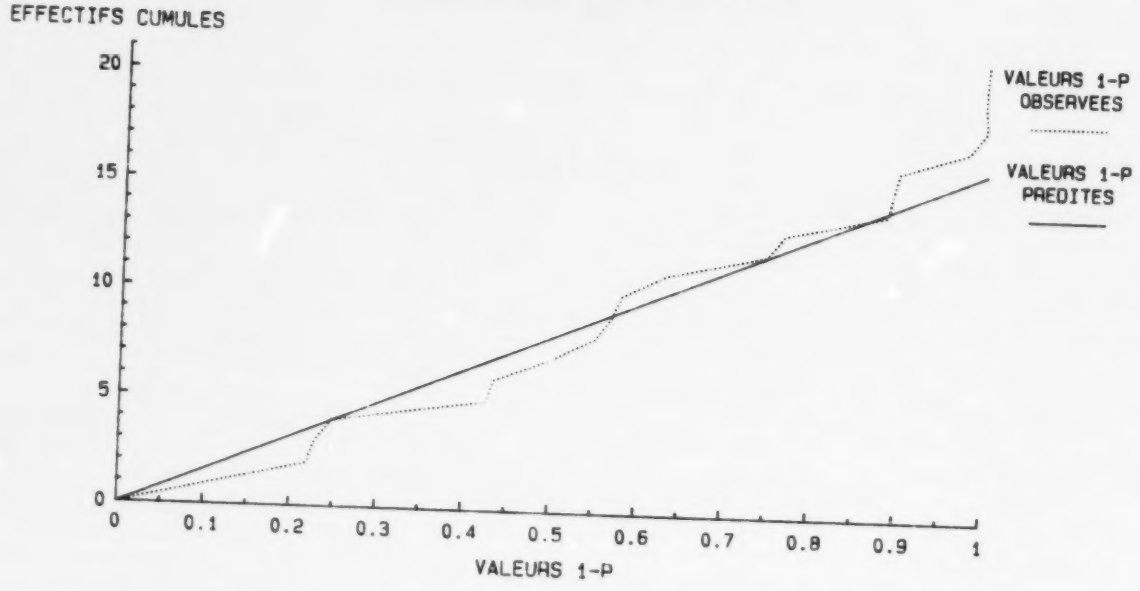
Un total de 21 tests ont été faits pour chacune des comparaisons entre sous-échantillons. Pour permettre le rejet, à un niveau de signification acceptable, d'hypothèses nulles, le seuil de signification doit être corrigé lors de l'utilisation simultanée de plusieurs tests. La correction de Bonferonni est connue pour son faible pouvoir. Des hypothèses nulles tendent à ne pas être rejetées lorsqu'elles le devraient. Ici, une procédure graphique a été utilisée pour estimer le nombre V d'hypothèses nulles vraies dans chacune des séries de 21 tests[15]. L'estimé V est utilisé pour obtenir un seuil de signification α tel que $\alpha = .05/V$.

RÉSULTATS

a) Les moyennes

Le graphe de la figure 1 a été utilisé pour estimer le nombre d'hypothèses nulles vraies dans le cas de la comparaison de la moyenne

FIGURE 1. GRAPHE-P DES TESTS T DES COMPARAISONS DES MOYENNES D'ATTITUDES



VALEUR APPROXIMATIVE DE LA PENTE: 16

des distributions de réponses aux 21 questions d'attitudes. La courbe des valeurs prédites a été ajustée par approximations successives comme dans les exemples donnés par Schweder et Spjotvoll(15). La valeur de la pente est de 16. Au seuil .05/16 = .003, les moyennes d'attitudes et de comportements ne diffèrent significativement que sur un seul couple de réponses. Les moyennes d'attitudes obtenues auprès du sous-échantillon qui a reçu deux questionnaires sont légèrement, mais non significativement, plus élevées que celles de l'autre sous-échantillon.

Les moyennes des distributions de réponses sur les questions relatives aux comportements diffèrent (tableau 1). En fait, la courbe des valeurs 1-p est si biaisée, qu'aucune pente n'a été obtenue dans ce cas. Le seuil de .05 a donc été utilisé. Des dix différences significatives entre les moyennes sur les questions se rapportant aux comportements, deux avaient des moyennes plus élevées dans le cas du sous-échantillon qui recevait un questionnaire. Généralement, les questions incluses dans les échelles EDU.1 et EDU.2 ont des moyennes plus petites pour l'échantillon ne recevant qu'un questionnaire que dans l'autre échantillon, tandis que le contraire se remarque pour les moyennes sur les questions incluses dans les échelles PSYCHO et INF. Ainsi, la cohérence des réponses aux questions sur les comportements et sur les attitudes semble dépendre du domaine de l'objet ou de l'acte envers lesquels les attitudes et comportements sont mesurés.

Tableau 1

MOYENNES DES COMPORTEMENTS

	Sous-échantillon		
	Un questionnaire	Deux questionnaires	Valeurs de P
ITEM2	4,925	5,036	,231
ITEM3	5,130	5,348	,016*
ITEM4	4,917	5,016	,161
ITEM5	4,934	5,078	,073
ITEM7	5,338	5,497	,033*
ITEM8	5,077	5,162	,144
ITEM9	4,838	4,979	,010*
ITEM17	2,870	2,802	,403
ITEM18	2,810	2,594	,003*
ITEM19	2,325	2,118	,002*
ITEM20	3,226	3,135	,463
ITEM21	2,490	2,374	,417
ITEM23	4,633	4,388	,299
ITEM26	4,518	4,262	,114
ITEM27	3,682	3,605	,364
ITEM29	4,396	4,933	,000*
ITEM30	4,009	4,635	,000*
ITEM34	4,794	5,070	,015*
ITEM39	4,150	4,335	,012*
ITEM40	4,448	4,755	,005*
ITEM41	4,204	4,771	,158

* Seuil de signification: α = ,05

b) Les variances

Les ratios de variance ont été calculés pour les réponses aux questions relatives aux attitudes dans chacun des échantillons. Les valeurs des ratios vont de 1,03 à 1,59. Seuls six ratios sont plus grands que 1,30. Le nombre d'individus incluent dans les échantillons sont suffisamment importants pour rendre tout ratio supérieur à l'unité statistiquement significatif. Ainsi, les ratios de variance n'ont qu'une valeur indicative des différences absolues entre les variances des deux sous-échantillons.

Les ratios de variance pour les mesures des comportements vont de 1,04 à 1,54. Seuls quatre ratios ont une valeur plus grande de 1,30.

La plupart des ratios de variance ont donc une valeur relativement petite ou moyenne. Aussi, les variances du sous-échantillon qui n'a reçu qu'un questionnaire sont plus élevées que les variances de l'autre sous-échantillon.

c) Les différences entre attitudes et comportements

Le nombre d'hypothèses nulles vraies a été établi à 8 selon la procédure graphique de Schweder et Spjotvoll(15). Cinq différences sont significatives au seuil de ,006. Ces différences confirment que les configurations de réponses aux questions d'attitudes et de comportements dépendent des domaines investigués. En effet, les différences entre attitudes et comportements correspondants sont plus importantes dans les domaines PSYCHO et INF pour le sous-échantillon qui recevait deux questionnaires, tandis que le contraire se remarque dans les domaines EDU.1 et EDU.2 (tableau 2).

Tableau 2

DIFFÉRENCES MOYENNES ENTRE ATTITUDES ET COMPORTEMENTS

	Sous-échantillon		
	Un questionnaire	Deux questionnaires	Valeurs de p
ITEM2	,259	,375	,209
ITEM3	,344	,185	,067
ITEM4	,392	,132	,002*
ITEM5	,502	,406	,214
ITEM7	,291	,163	,054
ITEM8	,298	,238	,453
ITEM9	,271	,156	,327
ITEM17	,777	1,019	,034
ITEM18	,687	1,165	,000*
ITEM19	1,031	1,555	,000*
ITEM20	,848	1,012	,164
ITEM21	,931	1,296	,001*
ITEM23	,230	,324	,293
ITEM26	,488	,653	,123
ITEM27	,837	,844	,951
ITEM29	,413	,212	,045
ITEM30	,681	,355	,000*
ITEM34	,625	,460	,062
ITEM39	,445	,462	,062
ITEM40	,411	,326	,408
ITEM41	,608	,495	,248

* Seuil de signification: $\alpha = .05/8 = .000$

d) Les corrélations

Le nombre d'hypothèses nulles vraies a été fixé à quatre dans le cas des comparaisons entre les corrélations des distributions d'attitudes et de comportements correspondants. Au seuil ,013, un total de huit corrélations sont significativement différentes d'un échantillon à l'autre (tableau 3). Les estimations des corrélations avec l'échantillon ne recevant qu'un seul questionnaire sont plus élevées que les estimations dans le cas de l'autre échantillon.

Les matrices de variance-covariance ont été comparées entre sous-échantillons, pour vérifier si les différences entre corrélations

Tableau 3

CORRÉLATIONS DES ATTITUDES AVEC LES COMPORTEMENTS

	Sous-échantillon		
	Un questionnaire	Deux questionnaires	Valeurs de p
ITEM2	,502	,352	,039
ITEM3	,349	,274	,352
ITEM4	,677	,329	,000*
ITEM5	,571	,382	,006*
ITEM7	,699	,536	,002*
ITEM8	,538	,576	,540
ITEM9	,594	,482	,028
ITEM17	,533	,352	,011
ITEM18	,416	,373	,566
ITEM19	,366	,317	,546
ITEM20	,407	,223	,025
ITEM21	,410	,339	,340
ITEM23	,693	,468	,000*
ITEM26	,470	,386	,248
ITEM27	,645	,446	,001*
ITEM29	,643	,404	,000*
ITEM30	,691	,625	,193
ITEM34	,662	,575	,117
ITEM39	,655	,436	,000*
ITEM40	,642	,365	,000*
ITEM41	,472	,319	,041

* Seuil de signification: $\alpha = .05/6 = .008$

pouvaient être attribuées aux différences déjà identifiées entre variances. Les comparaisons ont été faites en employant la procédure de Joreskog[16]. Les hypothèses d'égalité des matrices de variance-covariance ont été rejetées dans 18 cas sur 21. Mais l'hypothèse d'égalité des covariances n'a été rejetée que dans quatre cas uniquement. En conséquence, les différences entre variances ont un effet important sur l'estimation des corrélations.

CONCLUSION

Deux façons de procéder à la cueillette de données ont été évaluées. Dans un premier cas, des questions sur les attitudes et les comportements ont été intégrées à un seul questionnaire, dans un second cas, les biais systématiques dus à une cohérence dans les réponses aux questions relatives aux attitudes et aux comportements ont été contrôlés par l'emploi de deux questionnaires. Cette procédure double de cueillette de données a permis une comparaison des moyennes et des variances des attitudes, une comparaison des différences moyennes entre attitudes et comportements et une comparaison des corrélations entre attitudes et comportements.

Les comparaisons testées dans cette étude sont susceptibles d'un biais: un vrai changement a pu se produire entre le moment où les comportements et celui où les attitudes ont été mesurés auprès de l'échantillon de médecins qui ont reçu deux questionnaires. Le fait de répondre à un premier questionnaire donne aux médecins un temps de réflexion utile pour réfléchir sur les pratiques cliniques et modifier les idées et comportements. Aussi, des événements extérieurs peuvent se produire qui influencent le jugement du médecin sur sa pratique clinique. Dans notre étude, une période de deux semaines s'est écoulée entre l'envoi du premier questionnaire par le médecin et la réception du second, une période relativemnt courte. De même, les attitudes et les comportements mesurés concernent des dimensions importantes de la pratique médicale qui devraient être relativement stables dans le temps.

L'étude de la comparaison des moyennes d'attitudes et de comportements a démontré que les moyennes d'attitudes sont très semblables tandis que les moyennes de comportements diffèrent selon la procédure de collecte de données. Mais les moyennes des distributions des réponses sur les questions relatives aux comportements ne sont pas toutes plus élevées dans le sous-échantillon avec un questionnaire par rapport aux moyennes sur l'autre échantillon. Les différences sont généralement plus grandes dans le cas des dimensions orientées vers des questions rattachées à la pratique médicale usuelle, alors que le contraire a été observé pour les questions orientées vers un contenu psycho-social et informatif. Une configuration claire de biais systématiques apparaît ici. Elle est liée à la familiarité des répondants vis-à-vis les domaines d'attitudes et de comportements et, aussi, à leur difficulté à agir en fonction de leurs prédispositions. Donc, les biais systématiques introduits dans la mesure des attitudes et des comportements sont complexes puisqu'ils dépendent du type d'attitudes et de comportements mesurés.

Enfin, l'analyse des données montre clairement que les corrélations entre attitudes et comportements augmentent de façon systématique lorsqu'un seul questionnaire est utilisé. En effet, toutes les corrélations sont plus élevées dans le sous-échantillon qui ne reçoit qu'un seul questionnaire. Quoique certaines de ces différences soient attribuables aux variances, les relations entre attitudes et comportements sont habituellement estimées à l'aide des coefficients de corrélation. La mesure de ces relations est donc biaisée lorsqu'un seul questionnaire est utilisé.

En résumé, nous recommandons l'utilisation de deux questionnaires lorsqu'attitudes et comportements font l'objet d'une recherche.

RÉFÉRENCES

1. Ajzen I, Fishbein M: Attitudinal and Normative Variables as Predictors of Specific Behaviors. Journal of Personality and Social Psychology 1973;27:41-57.

2. Ajzen I, Fishbein M: Attitude-Behavior Relations: A Theoretical Analysis and Review of Empirical Research. Psychological Bulletin 1977;84:888-918.

3. Davidson AR, Jaccard JJ: Variables that Moderate the Attitude-Behavior Relation: Results of a Longitudinal Survey. Journal of Personality and Social Psychology 1979;37:1364-1376.

4. DeFleur ML, Westie FR: Attitude as a Scientific Concept. Social Forces 1963;42:17-31.

5. Fazco RH, Zanna MP: Attitudinal Qualities Relating to the Strenght of the Attitude-Behavior Relatioship. Journal of Experimental Social Psychology 1978;14:398-408.

6. Fishbein M: Factors Influencing Health Behaviors: An Analysis Based on a Theory of Reasoned Action. In F. Landy Health Risk Estimation, Risk Reduction and Health Promotion. Proceedings of the 19th Annual Meeting, the Society of Prospective Medicine, Québec, 1984:205-214.

7. Fishbein M, Jaccard JJ: Theoretical and Methodological Considerations in the Prediction of Family Planning Intentions and Behavior. Representative Research in Social Psychology 1973;4:37-51.

8. Heberlein TA, Black JS: Attitudinal Specificity and the Prediction of Behavior in a Field Study. Journal of Personality and Social Psychology 1976;33:474-479.

9. Hill RJ: Attitudes and Behaviors. In M. Rosenberg et RH Turner Social Psychology, Sociological Perspective. New York: Basic Books, 1981:347-377.

10. Liska AE: Emergent Issues in the Attitude-Behavior Consistency Controversy. ASR 1974;39:261-272.

11. Schuman H, Johnson HP: Attitudes and Behaviors. Annual Review of Sociology 1976;2:161-207.

12. Weigel RH, Newman LS: Increasing Attitude-Behavior Correspondance by Broadening the Scope of the Behavioral Research. Journal of Personality and Social Psychology 1976;33:793-802.

13. Weigel, Vernon DTA, Tognacci LN: Specificity of the Attitude as a Determinant of Attitude-Behavior Congruence. Journal of Personality and Social Psychology 1974;30:724-728.

14. Snedecor GW, Cochran WG: Statistical Methods, Iowa State University Press, 1967, p. 186.

15. Schewer T, Spjotvoll E: Plots of P-values to Evaluate Many Tests Simultaneously. Biometrika 1982;69:493-502.

16. Joreskog KG: Analysis of Covariance Structures. Scandinavian Journal of Statistics 1981;8:65-92.

Annexe

LES 21 QUESTIONS INCLUSES DANS LES QUATRE DIMENSIONS FACTORIELLES

Dimensions Items		Questions
	Q1	Le médecin omnipraticien est souvent appelé à suivre des patients souffrant de pathologies chroniques. Lorsqu'un médecin suit un hypertendu, un diabétique ou un arthritique par exemple, d'après vous, devrait-il renseigner le patient sur les aspects suivants de sa maladie?
EDU.1	ITEM2	Le nom médical de la maladie
EDU.1	ITEM3	La nature de la maladie (ex. chronicité, organe malade)
EDU.1	ITEM4	La cause de la maladie
EDU.1	ITEM5	L'évolution probable de la maladie, si traitée
EDU.1	ITEM7	Le nom du (des) médicament(s) prescrit(s)
EDU.1	ITEM8	Le type de médicament(s) prescrit(s) (ex. diurétique, hypotenseur)
EDU.1	ITEM9	Le rôle respectif de chaque médicament
	Q3	Toujours en vous référant aux mêmes types de patients, le médecin devrait-il faire les choses suivantes?
INF.	ITEM17	Ecrire les principaux renseignements sur une feuille comme aide-mémoire pour le patient
INF.	ITEM18	Donner au patient de la documentation écrite sur sa maladie
INF.	ITEM19	Reviser avec le patient lors de la visite suivante, la documentation qui lui a été donnée
INF.	ITEM20	Donner au patient les noms de ressources à contacter pour informations supplémentaires sur sa maladie (ex. diététiste, association, etc.)
INF.	ITEM21	Informer le patient sur les bons livres populaires publiés sur son problème de santé
PSYCHO	ITEM23	Chercher à connaître le milieu de vie du patient
PSYCHO	ITEM26	S'enquérir des inquiétudes du patient vis-à-vis de sa maladie
PSYCHO	ITEM27	Evaluer le support que le patient reçoit de ses proches

	Q5	Prenons maintenant un autre type de situation fréquemment rencontrée en pratique. Dans le cas où un jeune fumeur consulte au bureau pour un problème particulier non relié au tabagisme, croyez-vous que le médecin devrait faire les choses suivantes?
EDU.2	ITEM29	Obtenir des informations sur les habitudes tabagiques du patient
EDU.2	ITEM30	Vérifier si le patient connaît les risques du tabagisme pour sa santé
EDU.2	ITEM34	L'encourager à cesser de fumer
	Q6	Finalement dans le cas où un jeune obèse consulte au bureau pour un problème particulier non relié à son obésité, croyez-vous que le médecin devrait faire les choses suivantes?
EDU.2	ITEM39	Indiquer au patient quel devrait être pour lui le poids souhaitable
EDU.2	ITEM40	Lui offrir votre aide s'il décide de perdre du poids
EDU.2	ITEM41	L'encourager à maigrir

Analyse graphique des tableaux de fréquences.

JULIE BERNIER

Je vous exposerai dans les prochaines pages une méthode graphique pour simplifier l'analyse des tableaux de fréquences.

Lorsqu,on analyse un tableau de fréquences en ajustant un modèle loglinéaire, si le modèle choisi en est un avec interactions d'ordre supérieur, on est contraint à l'interprétation d'une grande quantité de paramètres μ;c'est là qu'une méthode graphique basée sur les méthodes traditionnelles d'analyse des données peut s'avérer utile.

La méthode est construite à partir d'un paramètre différent des paramètres μ et que j'etudierai par l'intermédiaire d'une statistique λ.

Soit un tableau de fréquences à trois variables XYZ avec X qui est la variable dépendante, et I,J, K le nombre de modalités de ces variables, formons

$$W = YZ$$

la combinaison des variables Y et Z ayant $J \times K$ modalités, étudier la relation entre X et Y, Z c'est étudier la relation entre X et W.Formons la statistique λ comme suit:

$$\lambda_{il}^{XW} = lnf_{il} - \sum_{i'} \frac{m_{i'.}}{m_{..}} lnf_{i'l} - \sum_{l'} \frac{m_{.l'}}{m_{..}} lnf_{il'} + \sum_{i'} \sum_{l'} \frac{m_{i'.} m_{.l'}}{m_{..}^2} lnf_{i'l'}$$

En fait, chaque indice l représente un couple (j,k) et on peut donc ecrire λ sous la forme:

$$\lambda_{il}^{XW} = lnf_{ijk} - \sum_{i'} \frac{m_{i..}}{m_{...}} lnf_{i'jk} - \sum_{j'} \sum_{k'} \frac{m_{.j'k'}}{m_{...}} lnf_{ij'k'} + \sum_{i'} \sum_{j'} \sum_{k'} \frac{m_{.j'k'} m_{i'..}}{m_{...}^2} lnf_{i'j'k'}$$

Voyons ce que cette statistique λ estime.

Soit $\{\mu, \mu^X, \mu^Y, \ldots, \mu^{XYZ}\}$ la paramétrisation loglinéaire des valeurs prédites, alors la quantité estimée par λ_{il}^{XW} s'exprime comme suit:

$$\psi_{ijk} = \gamma_{ijk}^{XYZ} + \gamma_{ij}^{XY} + \gamma_{ik}^{XZ}$$

où :

$$\gamma_{ijk}^{XYZ} = \mu_{ijk}^{XYZ} - \sum_{i'} \frac{m_{i'..}}{m_{...}} \mu_{ijk}^{XYZ} - \sum_{j'} \sum_{k'} \frac{m_{.j'k'}}{m_{...}} \mu_{ij'k'}^{XYZ} + \sum_{i'} \sum_{j'} \sum_{k'} \frac{m_{i'..} m_{.j'k'}}{m_{...}^2} \mu_{i'j'k'}^{XYZ}$$

$$\gamma_{ij}^{XY} = \mu_{ij}^{XY} - \sum_{i'} \frac{m_{i'..}}{m_{...}} \mu_{ij}^{XY} - \sum_{j'} \frac{m_{.j'.}}{m_{...}} \mu_{ij}^{XY} + \sum_{i'} \sum_{j'} \frac{m_{i'..} m_{.j'.}}{m_{...}^2} \mu_{ij}^{XY}$$

$$\gamma_{ik}^{XZ} = \mu_{ik}^{XZ} - \sum_{i,} \frac{m_{i'..}}{m_{...}} \mu_{i'k}^{XZ} - \sum_{k'} \frac{m_{..k'}}{m_{...}} \mu_{ik'}^{XZ} + \sum_{i'} \sum_{k'} \frac{m_{i'..} m_{..k'}}{m_{...}^2} \mu_{i'k'}^{XZ}.$$

C'est-à-dire que γ_{ijk}, γ_{ij}, et γ_{ik} sont respectivement des combinaisons linéaires des paramètres μ_{ijk} ,μ_{ij} et μ_{ik}.

Comment utilise-t-on cette décomposition du paramètre ψ_{ijk}? Et bien on peut montrer que :

$$H_0 : \gamma_{ijk}^{XYZ} = 0 \quad \forall i,j,k \quad \Longleftrightarrow \quad H_0 : \mu_{ijk}^{XYZ} = 0 \quad \forall i,j,k.$$

Et donc,sous H_0 on aura:

$$\psi_{ijk} = \gamma_{ij}^{XY} + \gamma_{ik}^{XZ}.$$

Les ψ_{ijk} se représentent sous forme d'un tableau de dimension $(J \times K) \times I$. Chaque ligne de ce tableau sera indicée d'un couple(j,k), chaque colonne d'une coordonnée i. Soit Ψ_{jk} la ligne en position (j,k), sous H_0:

$$\Psi'_{jk} = \begin{pmatrix} \gamma_{1j}^{XY} + \gamma_{1k}^{XZ} \\ \vdots \\ \gamma_{Ij}^{XY} + \gamma_{Ik}^{XZ} \end{pmatrix}.$$

Si je choisi une deuxième ligne ayant la même coordonnée j:

$$\Psi'_{jk'} = \begin{pmatrix} \gamma_{1j}^{XY} + \gamma_{1k'}^{XZ} \\ \vdots \\ \gamma_{Ij}^{XY} + \gamma_{Ik'}^{XZ} \end{pmatrix}.$$

Et ces deux vecteurs auront la propriété suivante:

$$\Psi'_{jk} - \Psi'_{jk'} = \begin{pmatrix} \gamma_{1k}^{XZ} - \gamma_{1k'}^{XZ} \\ \vdots \\ \gamma_{Ik}^{XZ} - \gamma_{Ik'}^{XZ} \end{pmatrix}.$$

C'est-à-dire que $\Psi'_{jk} - \Psi'_{jk'}$ est un vecteur qui ne dépend pas de j.Et donc la série de vecteurs $\Psi_{11}, \ldots, \Psi_{j1}$ est une translation de la série de vecteurs $\Psi_{12}, \ldots, \Psi_{j2}$.

Ceci signifie aussi qu'on a les équivalences suivantes:

$$\mu_{ijk}^{XYZ} = 0 \quad \forall i,j,k \quad \Longleftrightarrow \quad \gamma_{ijk}^{XYZ} = 0 \quad \forall i,j,k \quad \Longleftrightarrow \text{L'énoncé 1 est vérifié.}$$

ENONCÉ 1. *Dans $\mathbb{R}^I$, les lignes brisées joignant les modalités de Y pour une modalité fixée de Z sont parallèles.*

Mais ceci veut aussi dire que la projection de ces lignes brisées dans le plan des deux premiers axes principaux du tableau des λ_{il} devraient aussi respectér ce parallélisme.

C'est là qu'intervient la méthode graphique qui consiste à faire l'analyse des correspondances sur le tableau des λ_{il} afin de vérifier à quels endroits l'hypothèse du parallélisme est violée.

Voyons cette méthode à l'aide d'un exemple:

Page A1 : Tableau de fréquences.

Page A2 : Représentation par l'analyse des correspondances du tableau des λ_{ik}.

Page A3 : On relie les modalités de la variable ville pour chacun des syndicats. On note que le syndicat FTQ semble avoir une ligne brisée non parallèle aux autres.Donc si les μ_{ijk}^{XYZ} ne sont pas tous nuls , i.e. s'il y a interaction triple, ce pourrait être ce syndicat qui n'est pas conforme aux autres.Voyons ce que l'analyse par les modèles loglinéaires révèle:

Page A4 : Pour l'ensemble du tableau on rejette l'hypothèse d'une interaction triple nulle (p=.0298).Si par ailleurs on retire le syndicat FTQ du tableau de fréquences, on ne rejette plus cette hypothése (p=.7589).

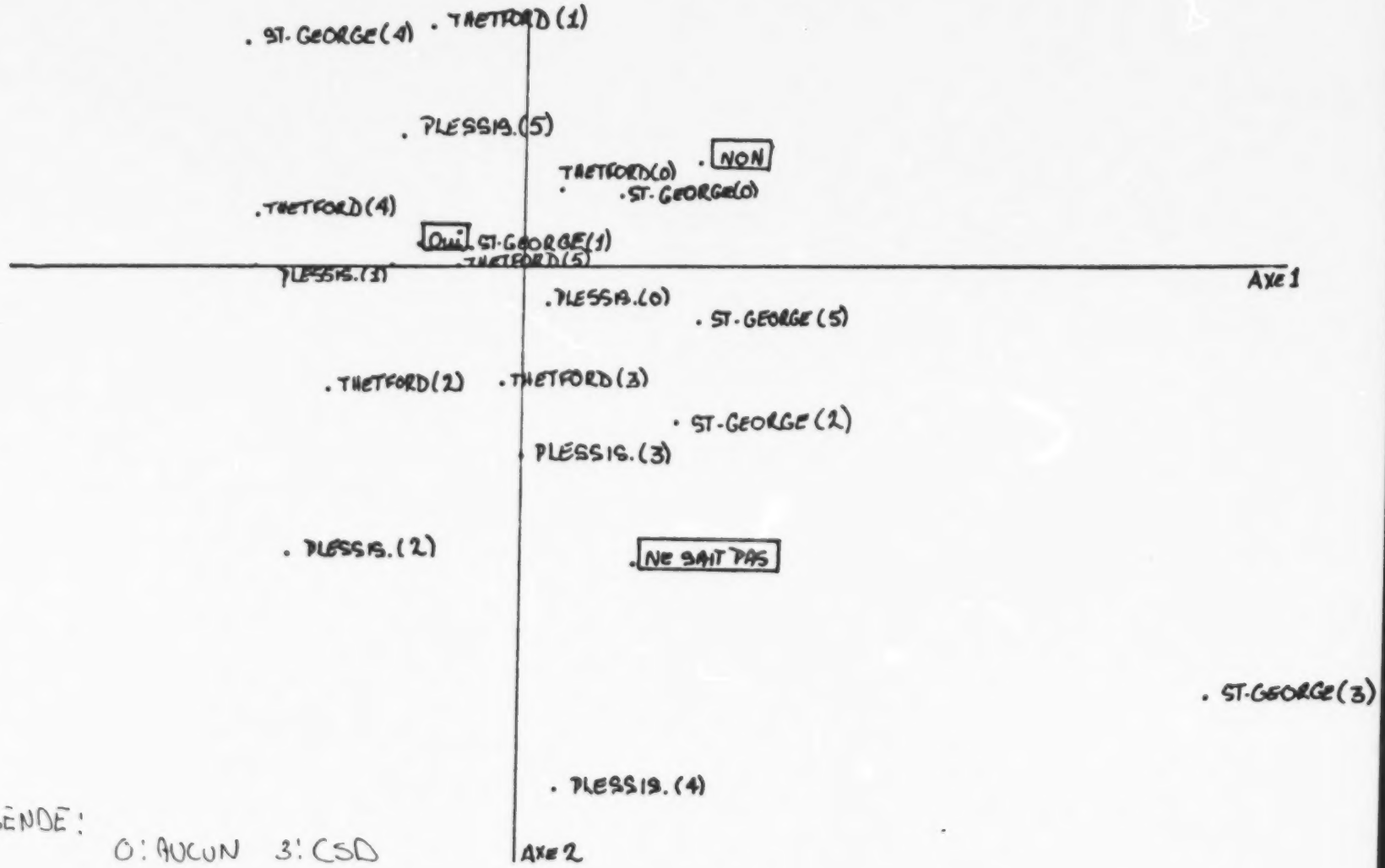
En fait, le modèle choisi serait celui où l'on conserve l'interaction entre ville et syndicat (p=.0000).La réponse à la question serait sans interaction avec la ville (p=.0684) et sans interaction avec le syndicat (p=.0726) et il faudrait interpréter séparément les réponses des membres de la FTQ.

A1- TABLEAU DE FREQUENCES DES REPONSES À LA QUESTION:

"Etes-vous pour ou contre les entreprises dirigées par les travailleurs?"

SYNDICAT	VILLE	REPONSE 0	1	2
AUCUN	PLESSISVILLE	18	52	24
	ST-GEORGE	23	69	49
	THETFORD	14	52	28
CEQ	PLESSISVILLE	5	27	7
	ST-GEORGE	5	23	8
	THE	1	13	5
CSN	PLESSISVILLE	4	15	2
	ST-GEORGE	4	6	4
	THETFORD	7	32	6
CSD	PLESSISVILLE	12	27	10
	ST-GEORGE	2	0	2
	THETFORD	0	1	0
FTQ	PLESSISVILLE	5	5	0
	ST-GEORGE	0	30	5
	THETFORD	10	60	15
INDEPENDANT	PLESSISVILLE	1	10	3
	ST-GEORGE	6	11	9
	THETFORD	2	10	3

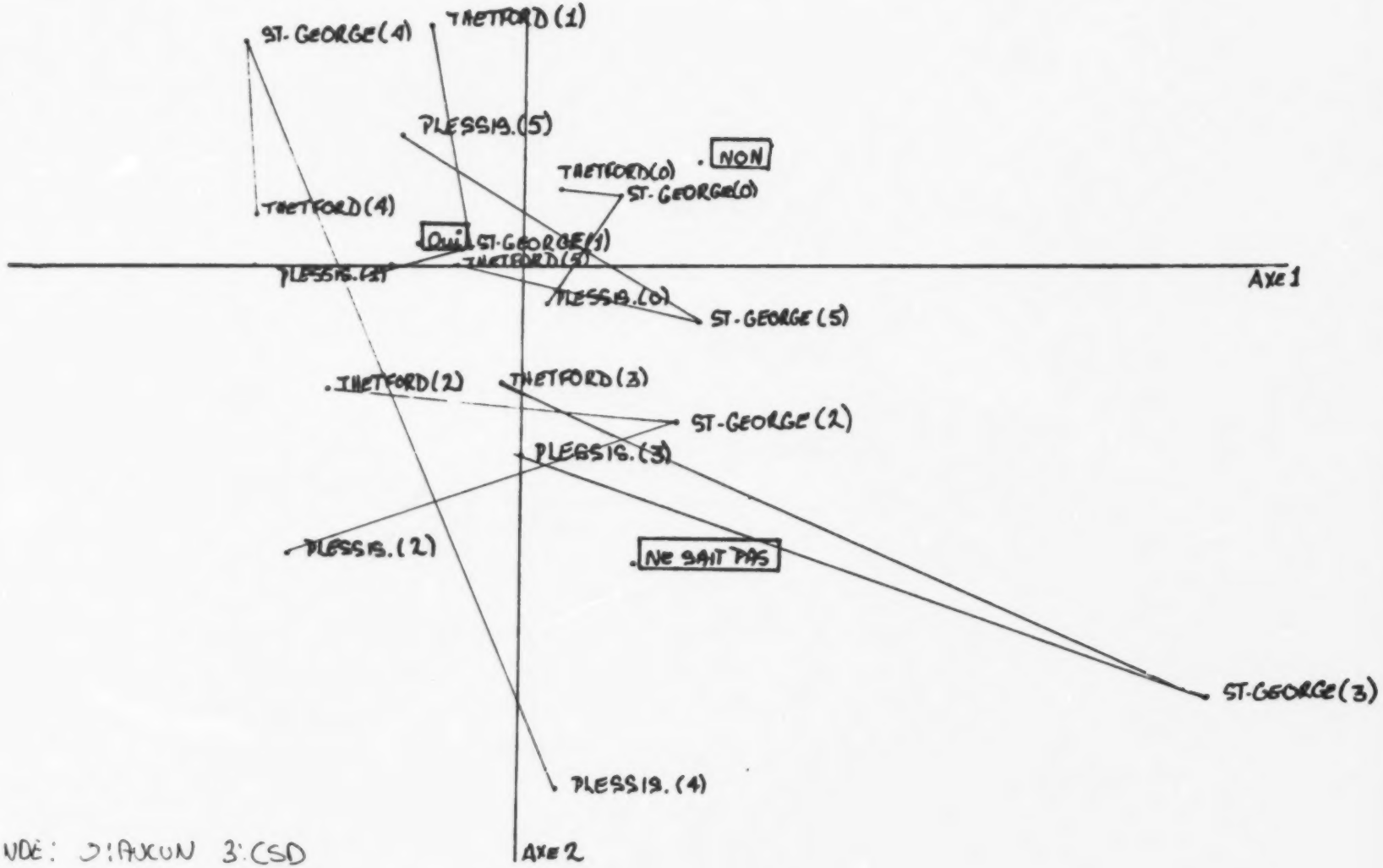
A2- REPRESENTATION DE LA VARIABLE W = Y*Z, LE PRODUIT DES DEUX VARIABLES INDEPENDANTES.



LÉGENDE:

0: AUCUN 3: CSD
1: CEQ 4: FTQ
2: CSN 5: INDEP.

A3- REPRESENTATION DE LA VARIABLE W = Y*Z, LE PRODUIT DES DEUX VARIABLES INDEPENDANTES.



LEGENDE: 0: AUCUN 3: CSD
1: CEQ 4: FTQ
2: CSN 5: INDEP

A4- RESULTATS DU "FIT ASSOCIATION" POUR L'ETUDE DU TABLEAU COMPLET.

EFFET	DEGRÉS DE LIBER.	CHI-DEUX	PROB.
R.	2	227.86	0.0
V.	2	2.56	0.2778
S.	5	356.93	0.0
RV.	4	7.06	0.1328
RS.	10	24.63	0.0061
VS.	10	188.42	0.0
RVS.	20	33.49	0.0298

RESULTATS DU "FIT ASSOCIATION" POUR L'ETUDE DU TABLEAU OBTENU APRES AVOIR ENLEVÉ LE SYNDICAT "FTQ".

EFFET	DEGRES DE LIBER	CHI-DEUX	PROB.
R.	2	153.84	0.0
V.	2	6.84	0.0328
S.	4	356.55	0.0
RV.	4	8.72	0.0684
RS.	8	14.37	0.0726
VS.	8	115.99	0.0000
RVS	16	11.78	0.7581

QUELQUES RÉFLEXIONS SUR L'INFÉRENCE STATISTIQUE DANS L'ANALYSE DE DONNÉES CATÉGORIELLES

LAMBERT Jean
Université de Montréal

1. INTRODUCTION

Il est de pratique courante d'utiliser une procédure asymptotique, par exemple le chi-deux de Pearson, afin de tester des hypothèses statistiques dans des tableaux de contingence. En se basant sur le théorème de la limite centrale, l'on s'attend à ce que ce genre de procédure soit approximativement valide en présence de grands échantillons. Toutefois, en présence de petits échantillons, l'on peut s'interroger sur la validité de telles procédures. Dans les lignes qui suivent, l'on tentera de répondre brièvement à cette question.

Toutefois, avant de répondre à cette question, on doit reconnaître que ce problème est intimement lié à la distribution de probabilités sous-jacente qui caractérise le tableau de contingence. Pour un tableau (r x c), on retrouve principalement trois situations telles que décrites dans le tableau I. En pratique, il est plus rare que l'on retrouve la situation I, à savoir, la seule situation n'occasionnant pas de difficulté majeure étant donné l'absence de paramètres de nuisance dans la distribution de probabilités sous-jacente. En effet, sans paramètres de nuisance, aucune estimation n'est requise dans l'utilisation d'une procédure asymptotique et celle-ci peut facilement se comparer au test exact basé sur la distribution hypergéométrique, mieux connu sous le nom de "test exact de Fisher".

Dans les situations II et III il est par contre plus difficile d'évaluer le bien-fondé d'une procédure asymptotique. En effet, en présence de paramètres de nuisance, une estimation est requise dans l'utilisation

d'une procédure asymptotique, et il n'existe pas de test exact. On peut tout de même investiguer la qualité d'une procédure asymptotique en générant empiriquement des tableaux de contingence à partir d'un modèle probabiliste pré-établi.

Selon les buts recherchés dans l'évaluation d'une procédure asymptotique l'on s'attardera à l'adéquation entre des degrés de signification exact et asymptotique, ou à l'ajustement d'une statistique de test à sa distribution d'échantillonnage théorique (principalement dans le voisinage des percentiles utiles aux tests d'hypothèse).

2. CRITÈRES D'ADMISSIBILITÉ DU X^2 DE PEARSON

En présence de petits échantillons, plusieurs règles empiriques ou subjectives ont été définies dans le but d'établir certaines normes ou critères au dessous desquels l'utilisation des procédures asymptotiques serait hasardeuse. Ces règles reposent en général sur:

- Les fréquences attendues minimums ou moyennes
- Les fréquences observées minimums
- La taille du (ou des) échantillon(s)
- Le seuil de signification
- La distribution de probabilités sous-jacente.

2.1 Situation hypergéométrique

Probablement le premier auteur à publier sur le sujet fut Yates (1934). A l'aide de quelques exemples numériques et en se basant sur sa vaste expérience, Yates suggéra qu'une fréquence attendue minimum (FAM) de cinq était nécessaire afin d'obtenir une bonne adéquation entre les degrés de signification du X^2 de Pearson et du test hypergéométrique exact dans le cadre d'un tableau de contingence (2 x 2), tableau II. Une FAM de un seulement était nécessaire s'il s'agissait du chi-deux corrigé avec la correction de continuité de Yates (dénoté X^2_c). Il est important de souligner que la correction de continuité de Yates, utilisée dans un tableau de contingence (2 x 2), n'est appropriée que dans la situation I puisque son but est de mieux approximer une probabilité hypergéométrique

(phénomène discret) à l'aide du X^2 (phénomène continu). Pour des tableaux de contingence avec plus d'un degré de liberté, Yates argumenta qu'il n'était pas nécessaire d'émettre des restrictions. Cependant, durant les années qui suivirent, et même récemment, la tendance est d'utiliser une FAM de cinq et même de dix (Hays, 1973) quels que soient les degrés de liberté et la statistique utilisée.

En plus de la FAM, Cochran (1952, 1954) introduisit la taille de l'échantillon comme critère d'admissibilité pour un tableau (2 x 2) afin de rendre moins sévères les normes sur la FAM. Mais il a surtout assouplit les normes d'une FAM de cinq ou de dix qui étaient de rigueur à l'époque pour des tableaux de contingence à plus d'un degré de liberté (tableau II).

Lewontin et Felsenstein (1965) s'intéressèrent plus à l'ajustement de la statistique X^2 à sa distribution théorique. Ils vérifièrent empiriquement q'une FAM de 0.5 était suffisante pour des tableaux de contingence ayant de 4 à 9 degrés de liberté en observant que le X^2 est une procédure très robuste en regard de l'erreur de 1ere espèce, tout en présentant un léger conservatisme (i.e., rejet de l'hypothèse nulle moins souvent que prévu au seuil de signification nominal), (tableau II).

Pour des tableaux de contingence (2 x 2) et en utilisant la fréquence observée minimum (FOM) comme critère d'admissibilité, Doane et Reese (1977) ont démontré que le X_c^2 est une procédure d'autant plus robuste que le seuil de signification augmente. Par exemple, pour un seuil nominal de 0.10, une FOM de zéro n'invalide pas le X^2_c (tableau II).

2.2 Situation d'un produit de multinomiales

Dans la situation II, Roscoe et Byars (1971) présentèrent des résultats élaborés, et détaillés en fonction des degrés de liberté, du seuil de signification et du niveau d'asymétrie de la distribution de probabilités sous-jacente (tableau III). Leur conclusion est que le X^2,

tout en présentant un léger conservatisme, est une procédure très robuste en regard de l'erreur de 1ere espèce, sauf dans le tableau de contingence (2 x 2) et dans des situations ayant plus d'un degré de liberté, une distribution de probabilités asymétrique et un seuil de signification bas.

Camilli et Hopkins (1978,1979), s'attardant au tableau de contingence (2 x 2) seulement, démontrèrent que le X^2 peut s'utiliser avec des effectifs d'échantillons aussi bas que deux ($n_1 = n_2 = 2$) ou un et trois ($n_1 = 1$; $n_2 = 3$) dans la mesure où le seuil de signification n'est pas trop bas. Dans ce cas, la procédure serait trop libérale (tableau III).

2.3 Situation d'un multinomiale générale

Dans la situation III, Craddock (1966) et Craddock et Flood (1970) ont étudié des tableaux de contingence avec plus d'un degré de liberté et leurs résultats démontrent qu'une fréquence attendue moyenne d'au moins un, était suffisante pour que le X^2 soit robuste (tableau IV). Ces résultats ont été en partie appuyés par Bradley et coll. (1979) qui ont insisté sur la symétrie de la distribution de probabilités sous-jacente.

Camilli et Hopkins (1978,1979) ont également étudié le tableau de contingence (2 x 2) dans la situation III. Leurs résultats ont démontré que le X^2 peut s'utiliser avec un échantillon d'effectif n=8 pour un seuil de signification de 0.05 ou 0.01 (et même avec un n=4 si le seuil de signification est de 0.10). Il y a risque d'un libéralisme en présence d'asymétrie (tableau IV).

3. RAPPORT DE VRAISEMBLANCE (G^2)

Le rapport de vraisemblance G^2 (re: Bishop et coll., 1975) n'a pas été autant étudié que le X^2 de Pearson. Cependant, on a démontré qu'il était très libéral par rapport à l'erreur de 1ere espèce, autant dans la situation I (Larntz, 1978) que dans la situation II (Grizzle, 1967; Odoroff, 1970; Margolin et Light, 1974).

4. CONCLUSION

A la lumière de ces résultats, le X^2 de Pearson peut être considéré comme la procédure la plus robuste si on le compare à ses compétiteurs asymptotiques, tels le rapport de vraisemblance, le X^2 de Neyman ou la procédure de Freeman-Tukey (Hutchinson, 1979). Quelle que soit la situation étudiée, on remarque que cette robustesse est associée à un léger conservatisme lorsqu'il y a plus d'un degré de liberté. Cependant, dans un tableau de contingence (2 x 2), les critères d'admissibilité sont plus sévères, et on note que le X^2 peut devenir libéral si ces critères ne sont pas respectés.

BIBLIOGRAPHIE

Bishop, Y. M. M., Fienberg, S. E. et Holland, P. W. (1975). Discrete Multivariate Analysis: Theory and Practice, The MIT Press, Cambridge, Massachusetts.

Bradley, D. R. et coll. (1979). "Type I Error Rate of the Chi-Square Test of Independence in RxC Tables that Have Small Expected Frequencies". Psychol. Bull., 86, 1290-1297.

Camilli, G. et Hopkins, K. D. (1978). "Applicability of Chi-Square to 2x2 Contingency Tables with Small Expected Cell Frequencies". Psych. Bull., 85, 163-167.

Camilli, G. et Hopkins, K. D. (1979). "Testing for Association in 2x2 Contingency Tables with Very Small Sample Sizes". Psych. Bull. 86, 1011-1014.

Cochran, W. G. (1952). "The X^2 Test of Goodness of Fit". Ann. Math. Statist., 23, 315-345.

Cochran, W.G. (1954). "Some Methods for Strengthening the Common X^2 Tests". Biometrics, 10, 417-451.

Craddock, J.M. (1966). "Testing the Significance of a 3x3 Contingency Table". The Statistician, 16, 87-94.

Craddock, J. M. et Flood, C. R. (1970). "The Distribution of the X^2 Statistic in Small Contingency Tables". Appl. Statist., 19, 173-181.

Doane, D. P. et Reese, R.M. (1977). "An assessment of the Accuracy of the X^2 Approximation to Hypergeometric Probabilities in 2x2 Contingency Tables Using Rules of Thumb". Proc. Stat. Comp. Sec. A.S.A., 185-189.

Grizzle, J. E. (1967). "Continuity Correction in the X^2-Test for 2x2 Tables". The American Statistician, 21, 28-32.

Hays, W.L. (1973). Statistics for the Social Sciences, 2nd ed., Holt, Rinehart and Winston, New-York.

Hutchinson, T. P. (1979). "The Validity of the Chi-Squared Test when Expected Frequencies Are Small: A List of Recent Research References" Commun. Stat. Theor. Meth., A8, 327-335.

Larntz, K. (1978). "Small-Sample Comparisons of Exact levels for Chi-Squared Goodness-of-Fit Statistics". J. Amer. Stat. Ass., 73, 253-263.

Lewontin, R. C. et Felsenstein, J. (1965). "The Robustness of Homogeneity Tests in 2xN Tables". Biometrics, 21, 19-33.

Margolin, B. H. et Light, R. J. (1974). "An Analysis of Variance for Categorical Data, II: Small Sample Comparisons with Chi Square and Other Competitors". J. Amer. Stat. Ass., 69, 755-764.

Odoroff, C. L. (1970). "A Comparison of Minimum Logit Chi-Square Estimation and Maximum Likelihood Estimation in 2x2x2 and 3x2x2 Contingency Tables: Tests for Interaction". J. Amer. Stat. Ass., 65, 1617-1631.

Roscoe, J. T. et Byars, J. A. (1971). An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic". J. Amer. Stat. Ass., 66, 755-759.

Yates, F. (1934). "Contingency Tables Involving Small Numbers and the X^2 Test". J. R. Stat. Soc., Suppl. 1, 217-235.

Odoroff, C. L. (1970). "A Comparison of Minimum Logit Chi-Square Estimation and Maximum Likelihood Estimation in 2x2x2 and 3x2x2 Contingency Tables: Tests for Interaction". J. Amer. Stat. Ass., 65, 1617-1631.

Roscoe, J. T. et Byars, J. A. (1971). An Investigation of the Restraints with Respect to Sample Size Commonly Imposed on the Use of the Chi-Square Statistic". J. Amer. Stat. Ass., 66, 755-759.

Yates, F. (1934). "Contingency Tables Involving Small Numbers and the X^2 Test". J. R. Stat. Soc., Suppl. 1, 217-235.

Tableau 1

Caractérisation d'un tableau de contingence (r x c) en fonction des distributions de probabilités sous-jacentes*

Caractéristique	Situation I	Situation II	Situation III
Distr. de probabilités:	Hypergéométrique (multivariée)	Produit de multinomiales	Multinomiale générale
Nombre de marges aléatoires:	Aucune	Une	Deux
Hypothèse testée :	"Distribution de A aléatoire p/r à B"	Test d'homogénéité $H_0: P_{1j}=P_{2j}=\ldots=P_{rj}$ (Pour tout j)	Test d'indépendance $H_0: P_{ij}=P_{i.}P_{.j}$ (Pour tout i,j)
Nombre de paramètres de nuisance:	Aucun	(c - 1)	(r + c - 2)
Le test exact existe-t-il?	Oui (test de Fisher)	Non	Non

* Ne sont présentées que les distributions de probabilités et les hypothèses les plus fréquemment utilisées.

Tableau II

Critères d'admissibilité du chi-deux de Pearson dans la situation hypergéométrique, avec (X^2_c) et sans (X^2) la correction de Yates

Degrés de liberté	Critère d'admissibilité	Commentaires	Procédure asymptotique
1. Yates (1934)			
1	$E(n_{ij})_{min} \geq 5$		X^2
1	$E(n_{ij})_{min} \geq 1$		X^2_c
> 1	(Suggère aucun)		X^2
2. Tendance après 1934			
≥ 1	$E(n_{ij})_{min} \geq 5$		X^2 , X^2_c
3. Cochran (1952, 1954)			
1	$E(n_{ij})_{min} \geq 5$	Si $20 \leq n \leq 40$	X^2_c
1	(Suggère aucun)	Si $n > 40$	X^2_c
> 1	$E(n_{ij})_{min} \geq 2$		X^2
> 1	$E(n_{ij})_{min} \geq 1$	Si au moins 80% des $E(n_{ij}) \geq 5$	X^2
4. Lewontin, Felsenstein(1965)			
4-9	$E(n_{ij})_{min} \geq 0.5$		X^2
5. Doane, Reese (1977)			
1	$n_{ijmin} \geq 0$	Pour $\alpha = 0.10$	X^2_c
1	$n_{ijmin} \geq 1$	Pour $\alpha = 0.05$	X^2_c
1	$n_{ijmin} \geq 5$	Pour $\alpha = 0.01$	X^2_c

Tableau III

Critères d'admissibilité du chi-deux de Pearson dans la situation d'un produit de multinomiales

Degrés de liberté	Critère d'admissibilité	Commentaires
1. Roscoe, Byars (1971)		
1	$E(n_{ij})_{moy} \geq 7.5$	
2-16	$E(n_{ij})_{moy} \geq 2$	Pour $\alpha = 0.05$, distr. sym.
2-16	$E(n_{ij})_{moy} \geq 4$	Pour $\alpha = 0.01$, distr. sym.
2-16	$E(n_{ij})_{moy} \geq 4$	Pour $\alpha = 0.05$, distr. mod. asym.
2-16	$E(n_{ij})_{moy} \geq 6$	Pour $\alpha = 0.01$, distr. mod. asym.
2-16	$E(n_{ij})_{moy} \geq 6$	Pour $\alpha = 0.05$, distr. très asym.
2-16	$E(n_{ij})_{moy} \geq 10$	Pour $\alpha = 0.01$, distr. très asym.
2. Camilli, Hopkins (1978, 1979)		
1	$n_1 = n_2 \geq 4$	Pour $\alpha = 0.01$
1	$n_1 \geq 2$ $(n_1/n_2 = 1/3)$	Pour $\alpha = 0.01$
1	$n_1 = n_2 \geq 4$	Pour $\alpha = 0.05$
1	$n_1 \geq 2$ $(n_1/n_2 = 1/3)$	Pour $\alpha = 0.05$
1	$n_1 = n_2 \geq 2$	Pour $\alpha = 0.10$
1	$n_1 \geq 1$ $(n_1/n_2 = 1/3)$	Pour $\alpha = 0.10$

Tableau IV

Critères d'admissibilité du chi-deux de Pearson dans la situation multinomiale générale

Degré de liberté	Critère d'admissibilité	Commentaires
1. Craddock (1966) Craddock, Flood (1970)		
2-16	$E(n_{ij})_{moy} \geq 1$	Pour $\alpha = 0.05$
2-16	$E(n_{ij})_{moy} \geq 1$	Pour $\alpha = 0.01$
2. Camilli, Hopkins (1978, 1979)		
1	$n \geq 8$	Pour $\alpha = 0.01$, sym.
1	$n \geq 8$	Pour $\alpha = 0.01$, 1 asym.
1	$n \geq 8$	Pour $\alpha = 0.01$, 2 asym.
1	$n \geq 8$	Pour $\alpha = 0.05$, sym.
1	$n \geq 8$	Pour $\alpha = 0.05$, 1 asym.
1	$n \geq 8$	Pour $\alpha = 0.05$, 2 asym.
1	$n \geq 4$	Pour $\alpha = 0.10$, sym.
1	$n \geq 4$	Pour $\alpha = 0.10$, 1 asym.
1	$n \geq 4$	Pour $\alpha = 0.10$, 2 asym.

PORTRAIT STATISTIQUE DE PEUPLEMENTS RESINEUX AYANT SUBI DES TRAITEMENTS SYLVICOLES

par
Mario Montégiani, Luc Perreault, Serge Simard

1. INTRODUCTION

Le cours "Projet", offert à l'Université Laval, offre aux étudiants la possibilité d'appliquer les diverses méthodes d'analyse apprises en classe à un domaine précis. C'est dans le cadre de ce cours que s'est déroulée l'étude décrite ici. Le Ministère de L'Energie et des Ressources nous a donc confié la tâche d'analyser différentes données provenant d'une étude relative à des travaux sylvicoles réalisés en Gaspésie de 1977 à 1983.

La section 2 permettra de se situer face à l'importance de la forêt au Québec tout en définissant ce que sont les travaux sylvicoles. La section 3 apportera les réponses aux questions suggérées par le Ministère alors que la section 4 répondra à des questions survenues lors de discussions avec les responsables de l'étude. Finalement, une conclusion sera présentée à la section 5.

2. IMPORTANCE DE LA FORET AU QUEBEC

Dans la vie de tous les jours, le bois et ses dérivés sont couramment utilisés par tout le monde. La grande versatilité du bois fait de cette ressource naturelle un secteur économique important au Québec. On peut d'ailleurs noter qu'un emploi sur dix y est relié.

Afin de conserver un niveau élevé de qualité des produits fores-

tiers, plusieurs moyens de conservation tels l'arrosage, le reboisement et les éclaircissements ont été instaurés. L'ensemble de toutes les actions entreprises dans ce sens se regroupent sous le terme de sylviculture.

Au cours des dernières années, le gouvernement du Québec a investi des sommes considérables dans la réalisation de travaux sylvicoles. C'est pourquoi une étude a été mise sur pied dans le but de vérifier l'état des arbres traités. Ceci permettra de connaître les résultats des travaux exécutés.

En plus de subir au moins un traitement sylvicole, les peuplements étudiés se sont vu accorder une protection supplémentaire soit la pulvérisation d'insecticide contre la tordeuse des bourgeons. Malgré tout, rien n'assure d'une qualité supérieure du bois. C'est pourquoi le Ministère de l'Energie et des Ressources a échantillonné près de 4400 arbres, mesurant sur chacun treize caractéristiques différentes.

3. REPONSES APPORTEES AUX QUESTIONS PROPOSEES PAR LE MINISTERE

Le principal but de notre travail était de répondre aux questions suivantes pour le bénéfice du Ministère:

1- Quels sont les éléments constituant l'échantillon?
2- Y a-t-il un lien entre les variables mesurées?
3- Quelles sont les caractéristiques de chaque essence?

La présente section traite à tour de rôle ces trois questions.

1- Quels sont les éléments constituant l'échantillon?

Aucun problème majeur n'est associé à cette question. Les statistiques descriptives sont tout-à-fait à propos pour permettre de caractériser et d'identifier les arbres de l'échantillon. Ainsi, à partir de nombreux tableaux et graphiques, nous avons été en mesure d'extraire les informations demandées dont

voici les plus importantes:

- La partie de la forêt sur laquelle ont été effectués les travaux sylvicoles peut se diviser en trois essences principales: l'épinette blanche (17%), l'épinette noire (11%) et le sapin beaumier (70%).

- Dans ces mêmes régions, les arbres sont très grands. En grande partie, ils appartiennent donc soit à la catégorie des dominants, soit à celle des co-dominants.

- La majorité des arbres étudiés sont situés à flanc de montagne ce qui signifie qu'ils sont exposés à des vents assez forts ainsi qu'à un drainage assez rapide.

2- Y a-t-il un lien entre les variables mesurées?

Le but visé par cette question était de pouvoir établir des liens entre les différentes variables pour ainsi pourvoir en réduire le nombre sans perdre trop d'information. Certains problèmes ont été rencontrés, dus au fait que les responsables au Ministère de l'Energie et des Ressources désiraient obtenir des liens bien particuliers mais conditionnés par d'autres variables. Le fait que les variables étaient toutes catégoriques avec peu de modalités a également compliqué l'analyse.

Afin de résoudre ce problème, nous avons décidé d'utiliser des tests de G^2 et de X^2 sur des tableaux de contingence pour l'indépendance des variables deux à deux et des modèles log-linéaires pour le reste. Ces méthodes nous ont permis de découvrir que, bien que toutes les variables soient significativement différentes deux à deux, l'information peut être résumée par cinq ou six variables seulement à l'aide des modèles log-linéaires.

3- Quelles sont les caractéristiques de chaque essence?

Le principal problème rencontré ici découlait du fait que nous avions à étudier plusieurs variables, certaines ayant jusqu'à cinq modalités. Nous avons donc dû utiliser une généralisation de l'analyse factorielle des correspondances, l'analyse factorielle des correspondances multiples, qui nous a permis de traiter toutes les variables et d'obtenir les caractéristiques de chaque essence.

Le point le plus important découlant de cette analyse, point très intéressant pour les chercheurs responsables de l'étude, est que le sapin beaumier est de loin le mieux adapté et le plus fort des bois tandis que l'épinette blanche est la plus fragile.

4. REPONSES AUX QUESTIONS SUSCITEES PAR L'ANALYSE

L'ensemble des questions provoquées par le déroulement de l'analyse peuvent se résumer dans celle-ci: "Quelle serait la meilleure façon de connaître l'évolution des peuplements?" Toutes ces questions découlaient en fait de la difficulté qu'avaient les chercheurs à déterminer des critères rigoureux et reconnus pour définir la santé d'un arbre.

Pour répondre à la question précédente, il nous apparaissait logique de, d'abord, construire un plan d'échantillonnage pour ensuite modéliser la santé de l'arbre.

i) Plan d'échantillonnage

Pour réaliser le plan d'échantillonnage, nous avons considéré les quelques 4400 arbres échantillonnés comme étant les individus constituant notre population. Le problème pourrait être formulé comme suit: "Quelle serait la manière la moins coûteuse

qui nous permettrait de connaître le pourcentage d'arbres en santé dans nos peuplements?"

Afin d'éviter les déplacements trop fréquents et pour assurer une bonne représentativité, nous avons choisi un plan d'échantillonnage à deux degrés. Par la suite, nous avons post-stratifié par la variable essence. Nous choisirons quatre régions au premier degré et trois zones dans chaque région (deuxième degré). On peut résumer la situation comme suit:

Degré	Probabilité de sélection
1	1/N (sans remise)
2	$1/M_i$ (sans remise)

Finalement, l'estimateur qui nous apparaît le plus adéquat pour ce plan d'échantillonnage est celui d'Horvitz-Thompson, défini comme suit:

$$\text{Proportion} = \bar{Y} = \frac{1}{n} \sum_i N W_i \bar{y} \qquad \text{où} \quad W_i = \frac{M_i}{\sum_i M_i}$$

ii) Modélisation de la santé de l'arbre

L'objectif principal de cette partie est de quantifier la santé de l'arbre. Pour ce faire, il fallait bâtir une variable qui permettrait de distinguer les arbres en santé de ceux qui ne le sont pas. Après de longues discussions avec les responsables de l'étude, nous sommes arrivés à la variable suivante (qui sera appelée ici la vulnérabilité de l'arbre):

une tête		plus d'une tête		
FTA=1	FTA=2	TS	TC ou TI	
non vulnérable	?	FTA=1	FTA=2	
		?	vulnérable	vulnérable

Les symboles utilisés doivent être interprétés comme suit:

FTA=1 : présence d'une flèche terminale
FTA=2 : absence d'une flèche terminale
TS : tiers supérieur
TC : tiers central
TI : tiers inférieur
? : impossible à classer

La variable réponse, la vulnérabilité, étant trouvée, il ne reste plus qu'à déterminer les variables explicatives pour construire notre modèle. Les variables choisies sont: le drainage, le diamètre à hauteur de poitrine, la défoliation totale, la pousse et le branchage.

Nous avons choisi, pour expliquer la santé de l'arbre, d'utiliser un modèle de régression logistique. Notre choix se base sur les deux raisons suivantes:

- ce type de modélisation correspondait exactement à notre objectif principal soit de quantifier la santé de l'arbre;
- un modèle de régression linéaire conventionnel ne pouvait s'appliquer.

La dernière raison peut s'expliquer ainsi. Prenons le cas général. Nous sommes en présence d'une variable réponse de type binaire.

Soit Y qui est égale à 0 ou 1 et $X_1,\ldots,X_n$ les variables explicatives. Posons que

$$P[Y_i=1] = \pi_i \quad \text{et ainsi} \quad \hat{Y} = E[Y_i] = [1 \times P[Y_i=1]]+[0 \times P[Y_i=0]] = \pi_i$$

La prédiction est donc une probabilité. Lorsque nous appliquons la méthode des moindres carrés ordinaires à ce cas, nous nous

heurtons à trois problèmes:

1- Il n'y a pas d'homogénéité des variances.
2- Le terme d'erreur n'obéit pas à la loi normale.
3- Il y a une contrainte sur la fonction de réponse due au fait que la valeur prédite est une probabilité.

Les remèdes proposés sont:

1- Pour palier à l'hétérocédasticité , nous avons appliquer les moindres carrés pondérés où la pondération sera l'inverse des variances de chacun des niveaux. La variance est ainsi stabilisée.
2- Nous savons que la méthode des moindres carrés donne des estimateurs asymptotiquement non biaisés et normaux. Ainsi, avec un grand échantillon comme c'est le cas ici, nous pouvions supposer la normalité du terme d'erreur.
3- Nous avons dû borner la fonction de réponse comme suit:

$$0 \quad E[Y_i] \quad 1$$

La fonction logistique est celle qui répond le mieux aux contraintes citées plus haut. Ainsi

$$E[Y_i] = \P_i = \frac{\exp(X_i)}{1+\exp(X_i)}$$

Par exemple, nous avons obtenu, de cette façon, le modèle suivant pour l'épinette noire:

$$P[V=1] = \frac{\exp(-0.1136 + 0.5003 \times DRAIN - 0.1611 \times POUSSE)}{1 + \exp(-0.1136 + 0.5003 \times DRAIN - 0.1611 \times POUSSE)}$$

où V = 0 si l'arbre est non vulnérable
1 si l'arbre est vulnérable

5. CONCLUSION

Avant de terminer, il est important d'attirer l'attention sur la subjectivité de nos données. En effet, pour la plupart des variables, l'évaluation du résultat a été faite à l'oeil. Evidemment, ceci affecte la précision des données. Nous suggérons, pour des études futures, de définir des variables qui réduiraient les erreurs dues à l'évaluation. Il serait préférable d'effectuer de véritables mesures qui donneraient des valeurs numériques quantitatives continues.

Enfin, il aurait été intéressant d'avoir à notre disposition un ensemble d'arbres témoins, non traités. Nous aurions alors pu comparer les arbres traités avec ceux non traités et répondre à des questions plus particulières comme, par exemple:

- Est-ce que les travaux sylvicoles améliorent les peuplements?
- Pour une certaine essence, plus résistante, est-il utile de faire des travaux sylvicoles?

Changement d'adresse

S'il-vous-plaît, remplir le présent avis et le retourner à:

Bureau de la statistique du Québec
Centre de distribution
117, rue Saint-André, 3e étage
Québec (Québec)
G1K 3Y3

Nom: _____ **Numéro d'abonné:** _____
(numéro inscrit sur l'étiquette)

Ancienne adresse:

_____
_____
_____
_____
_____

Nouvelle adresse:

_____
_____
_____
_____
_____